Inference for epidemics and effect of reporting processes

Kokouvi Mawuli Gamado

Submitted for the degree of
Doctor of Philosophy
on completion of research in the
Department of Actuarial Mathematics & Statistics,
School of Mathematical and Computer Sciences,
Heriot-Watt University

February 2012

I hereby declare that the work presented in this thesis was carried out by myself at Heriot-Watt University, Edinburgh, except where due acknowledgement is made, and has not been submitted for any other degree.

Kokouvi Mawuli Gamado (Candidate)

Dr George Streftaris (Supervisor)

Dr Stan Zachary (Supervisor)

Date

Abstract

The objective of this thesis is to study the effect of under-reporting in epidemics. In particular, there are two broad questions we investigate:

- In the situation of under-reporting in epidemics, what would happen if the data were treated as if no under-reporting were occurring? Such assumption leads to an under-estimation of the contact rate, implying an under-estimation of the reproduction number.

- By allowing for the fact that under-reporting is occurring, how and how well can we estimate the reporting rate and other parameters of the model?

We explore the above questions by considering the stochastic Markovian SIR epidemic in which various reporting processes are incorporated. We consider cases of constant reporting probability and move on to more realistic assumptions such as the reporting probability depending on time, the number of reported cases and the dependence on the source of infection for each infected individual.

We develop various methodologies, based on temporal data, to account for under-reporting in the Bayesian framework using MCMC to sample from the posterior distributions of the model parameters.

An introduction to the spatial aspect is also considered with the SIR model with reporting process on $\mathbb{Z}$.

Acknowledgements

I wish to express my deepest appreciation and gratitude to my supervisors Doctor Stan Zachary and Doctor George Streftaris for their guidance and encouragement. I have learnt to grow up a lot under your supervision not only in research but also in personal life.

My sincere thanks to Doctor Bernd Schroers for all his support, especially phoning me about the funding opportunity while I was still in South-Africa, and encouraging me to come and study at Heriot-Watt University.

I am especially grateful to Professor Gavin Gibson for all his support and advice throughout this project.

I would like to thank Professor Denis Mollison for all his advice.

I am grateful to the Actuarial Mathematics & Statistics department at Heriot-Watt University for providing funding.

To you my dad and mum, Essè and Anti Gamado, my brothers and sisters, uncles, cousins and nephews, I say thanks for your support and unwavering encouragement: your belief in me is my first source of motivation.

My sincere appreciation to all my friends and colleagues at Heriot-Watt University for the great moments we shared together.

My life in Edinburgh has been blessed by so many people that I cannot name all here. You all know who you are and I just want to say thanks for creating a family atmosphere around me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Modelling epidemics is still one of the great challenges faced by epidemiologists, mathematicians and statisticians. There exist a large variety of epidemic models that are widely applied. However, case-specific models are still needed to draw conclusions and take policy decisions for control and prediction of outbreaks. The general framework for studying disease evolution is to divide the population into compartments. Inference about the observed underlying process regularly assumes perfect reporting while this does not always happen to be the case. The recent pandemic of the influenza A (H1N1), in 2009, is an illustration where there was evidence of under-reporting, i.e. not all the infected cases were reported. When providing early findings for the pandemic H1N1, Fraser *et al.* (2009) considered under-reporting which they explicitly introduced in their modelling.

The question of under-reporting of infected cases is crucial for some diseases because of the bias it can introduce when making inference for the model parameters. We are interested in the effect of under-reporting in epidemics in this thesis. In particular, there are two broad questions we should like to investigate:

1. What happens if infected cases are not fully reported and we treat the data as if under-reporting is not occurring? We expect this to generally lead to under-estimation of infection rates, and therefore, of the reproduction number.

2. Assuming we make allowance for the fact that under-reporting is occurring, how, and how well, can we estimate the under-reporting and other parameters of the

model?

We consider the general stochastic epidemic defined in Subsection 1.3.1 in which we incorporate reporting processes. There are different variants of the reporting considered throughout this thesis: the reporting may happen immediately after either infection or removal. Models in which the reporting process is time-dependent and cases of the reporting depending on the source of infection are also considered.

1.2 Epidemics modelling

1.2.1 Reasons for modelling epidemics

When observing or after observation of an epidemic, it is crucial to provide insights of the underlying process in order to understand, control and predict outbreaks. However, statistical analysis of infectious disease is typically not straightforward, requiring the development of problem-specific methodology. The nature of epidemic data makes their statistical analysis not always easy.

The analysis of outbreak data can be more effective when it is based on the model for the actual dynamic process that generates the data. Moreover, models can be used to provide a better understanding of the infection process, the transmission dynamics and the epidemiologically important quantities of interest. There exist a number of reasons for modelling epidemics and making inference using historical incidence data. Analysis of this kind is used for diseases occurring due to novel or re-emerging pathogens. The review by Ferguson *et al.* (2003) gives a good description for models of historical incidence data. The threat of the recent highly pathogenic avian influenza disease in 2009 is an illustration. In 2003, the world also experienced a SARS outbreak (see for example Lipsitch *et al.* (2003) and Riley *et al.* (2003)), having significant impacts on public health. In 2001, the UK suffered a Foot-and-Mouth epidemic with significant economic impact as described by Bennett *et al.* (2001). Some deliberately released pathogens such as smallpox (see e.g. Kaplan *et al.* (2002)) remain a threat for populations. Ferguson *et al.* (2003) argue that there does not exist an epidemic model that can be "truly predictive" in the context of smallpox outbreak planning, and as a consequence no control method can be *a priori* identified as absolutely optimal. However a range of models and a set of control options are vital to know and need to

be adjusted in the event of an outbreak.

Models help to provide estimates of the parameters of interest that are responsible for driving the dynamics of the disease, and also answer some questions referring to the progress of the disease based on the current state of the outbreak. Epidemic models can, in addition, play a key role in control strategies, guiding the action of effective control policies to prevent a major spread. They can also be used to design optimal vaccination strategies. Despite the desire to design complex models that highlight every aspect of the disease and the population, it is also important to construct models for which inference can be drawn and interpretation of the parameters can be made to bridge the gap between modellers and policymakers.

1.2.2 Deterministic vs Stochastic

We are interested in statistical analysis of data and accordingly, our focus is on stochastic or probability models rather than deterministic ones. Disease propagation is an inherently stochastic phenomenon as justified below.

Real-life epidemics can either go extinct with an ultimate small number of infections, or end up with a significant proportion of the population having contracted the disease. In other words, there is a "chance" factor in the disease transmission from one individual to another. Therefore, there is a need for models to ascribe the unpredictable aspects of real epidemics to an element of chance, i.e. stochastic models. Moreover, stochastic models are intuitive since they naturally capture the infection process between different individuals. Deterministic models can also be fitted to data and thereby lead to estimates for parameters, but it can be difficult to assess the precision of such estimates. The natural role for the deterministic model is as an approximation to the stochastic model when all population sizes are large. This is the view in the past of deterministic models as claimed by Isham (2005). Deterministic models are also seen to be more useful in enriching the general theory of epidemic than in applications to real data. However, it is now widely accepted that both deterministic and stochastic models have their strengths and both contribute to good understanding of the underlying process (Renshaw, 1993). It is nevertheless becoming more apparent to scientists, biologists in particular, that models for epidemics need to incorporate intrinsic stochasticity in many ways. But the need for realistically

complex models has made deterministic models very popular and there is a need to develop general statistical methods to analyse complex stochastic models. In this thesis, we are focused on stochastic models and shall now describe the models in details, reviewing models commonly used in the literature.

1.3 Compartmental models for epidemics

Stochastic epidemics are often modelled describing the transitions of individuals through various disease-development states. Individuals move typically from a susceptible (S) to an infectious (I) state before their recovery or removal (R). The term *susceptible* means that at this stage, the individual has not contracted the disease but is not exempt of contracting it. *Infectious* refers to the fact that the individual has contracted the disease and at the same time is able to contaminate or pass on the disease to other susceptibles. The terms infectious and infected are the same in this context. An individual is considered as *removed* when he was infected and has recovered from the disease or becomes immunized or dead; in any case, he plays no further role in the spread of the disease. Additionally, more states can be required for the model such as the *exposure* state (E) , where an infected individual passes through a latent period before becoming infectious, or the *notified* (N) state (Jewell *et al.*, 2008), where infected individuals are reported and their infectiousness is reduced using control measures. Variations of the SIR and SEIR models can be found for instance in Bailey (1996) and Keeling and Rohani (2007). We are interested in the SIR model in particular in this thesis and present below the details of the transitions between states.

1.3.1 SIR Models in a homogeneous population

We consider a closed population (i.e. no births/deaths/immigration) of size N among which we assume that at time $t = 0$, there are a initially infected individuals. We denote by $S(t)$, $I(t)$ and $R(t)$, respectively the size of compartments of susceptible, infected and removed individuals in the population at time t. The assumption of

closed population implies that at each time t,

$$S(t) + I(t) + R(t) = N. \tag{1.1}$$

We also assume that the population is homogeneous meaning that all the individuals are considered to be of similar nature and homogeneously mixing.

Markovian SIR epidemic

The Markovian SIR model is characterised by the transition probabilities:

$$\Pr\left(S(t+dt) = S(t) - 1\right) = \beta S(t) I(t)\, dt + o(dt) \tag{1.2}$$

$$\Pr\left(I(t+dt) = I(t) - 1\right) = \gamma I(t)\, dt + o(dt) \tag{1.3}$$

The parameter β denotes the infection rate per susceptible-infected contact. It is assumed that every contact between susceptible and infected is potentially a disease transmission contact. Equation (1.2) implies that the infection times follow a time-varying Poisson process with rate $\beta S(t)I(t)$. The removal rate is γ and the process (1.3) means that the infectious period follows an Exp(γ) distribution. Due to the lack of memory property of the exponential distribution, the model is called Markovian as it can be fully described by continuous time Markov chains. The Markovian SIR epidemic is also referred to as the general stochastic epidemic and has been generalised in many ways. It is easy to simulate such an epidemic using the Gillespie algorithm (Gillespie, 1977).

Generalised stochastic epidemic

The assumption of exponentially distributed time to infection is unrealistic for many diseases. Indeed, the memoryless property of the infectious period supposes that the time an infected individual remains infectious is independent of the time he has already spent with this disease. However, for most diseases, the time an infected individual will still spend with the disease is influenced by the time he got infected. Therefore, we need a more flexible distribution for the infectious period.

The generalised stochastic epidemic model is the same as the general Markovian

epidemic except from that the infectious periods of different infectives are i.i.d. and assumed to follow a specified distribution which is not necessarily exponential. The most common distributions used in the literature for the disease lifetime are the Weibull distribution (Streftaris and Gibson, 2004a,b) and the Gamma distribution (O'Neill and Becker, 2001; Jewell *et al.*, 2008). The reason for the choice of the Weibull and Gamma distributions are due to their flexibility and capacity to mimic other statistical distributions. The particularity of this model now is that we need to consider individually each member of the population with its infection and removal time i.e, there is a need to pair infection and removal times of infected individuals.

Basic reproduction R_0 and the threshold result

A very important parameter in epidemic modelling is the basic reproduction number R_0 defined as the expected number of secondary infections generated by a single, typical infection in a completely susceptible population (see Heesterbeek and Dietz, 1996).

The Markovian SIR epidemic model described above assumes homogeneous population. A typical individual can then be any of the infectives and will, on average, be infectious for time $1/\gamma$. The number of susceptibles infected by one infective per unit time is $\beta(N-1)$. Therefore, the expected total number of infections produced by one infective is $R_0 = \beta(N-1)/\gamma$. In the case of the generalised stochastic epidemic, the basic reproduction ratio is defined as

$$R_0 = \beta(N-1)\boldsymbol{E}[\mathbb{I}] \tag{1.4}$$

where $\boldsymbol{E}[\mathbb{I}]$ is the expectation of the infectious life time random variable $\mathbb{I}$.

The definition of R_0 for more complicated models is not straightforward and one needs to be careful when defining an appropriate measure. R_0 is also called the *threshold parameter* since its value determines whether a "major" epidemic can occur or not. Indeed, the threshold theorem Whittle (1955) stipulates that when $R_0 \leq 1$, with probability one, only a finite number of individuals will become infected in an infinite population, i.e. the epidemic will die out. But when $R_0 > 1$, there is a positive probability that an infinitely large number of individuals will contract the disease in question. This suggests that to eradicate an epidemic, it is sufficient to

reduce the basic reproduction number to be less than one through different measures like vaccination, isolation barrier, etc.

The threshold theorem remains the most important result in the mathematical theory of epidemics since its introduction by Whittle (1955), see also Williams (1971) and Ball (1983). It is an asymptotic result and it is more difficult to define minor and major outbreaks for finite populations. However, it is broadly true that an epidemic will either very likely die out with minor impact or else might end up with a large proportion of susceptibles getting infected. This can be verified through simulations of such epidemics using the Gillespie algorithm.

Final size distribution

The final size of an epidemic is defined as the number of initially susceptible individuals that ultimately become infected. The distribution of the final size, for the generalised stochastic epidemic, can be obtained by solving a system of triangular equations that we need to introduce. Let $\phi(\theta) = \boldsymbol{E}\left[\exp(-\theta\mathbb{I})\right]$ be the moment generating function of the infectious period $\mathbb{I}$ and P_N^k the probability that the final size of the epidemic is equal to k where $0 \leq k \leq N$. Using coupling arguments, Ball (1986) (see also Mollison, 1995) showed that P_N^k satisfies the triangular system of equations:

$$\sum_{k=0}^{l} \frac{\binom{N-k}{l-k} P_N^k}{\left[\phi\left(\beta(N-l)\right)\right]^{k+a}} = \binom{N}{l}, \quad \text{for } l = 0, \ldots, N. \tag{1.5}$$

Due to numerical rounding errors, solutions of Equation (1.5) can lead to negative probability values. Demiris in his thesis (Demiris, 2004) discussed that even for moderate population sizes greater than 100, those numerical problems occur with certainty. To avoid the numerical problems, the approach proposed there (see also Demiris and O'Neill, 2005) is the multiple precision arithmetic which is computational costly and time consuming.

Gaussian approximation

Demiris (2004) used a Gaussian approximation to the final size distribution provided by Andersson and Britton (2000).

Assume that we have a sequence of Generalised Stochastic Epidemics indexed by

the initial susceptible population N. K denotes the final size of the N^{th} epidemic and let τ be the asymptotic proportion of individuals ultimately infected, i.e. $\tau = \lim_{N\to\infty} \frac{K}{N}$. τ is almost surely a constant and in the case $R_0 > 1$, τ is the non-trivial solution of the non-linear equation

$$1 - \tau = \exp(-R_0\tau). \tag{1.6}$$

It is easy to notice that 0 is always a solution of (1.6). Equation (1.6) is straightforward to prove using the equivalent of the stochastic Markovian SIR epidemic in the deterministic case (Kermack and McKendrick, 1927). A general proof for stochastic formulation of the model is given by Andersson and Britton (2000). A possible interpretation of this result can be provided by looking at the left and right hand sides separately. The probability of escaping infection, for an individual faced by τ attacks each affecting on average R_0, is equal to 0 occurrence for the Poisson(τR_0), i.e. the right hand of (1.6). On the other hand, the probability of escaping infection is equal to the proportion of initial susceptibles who remain uninfected, i.e the left hand side of (1.6).

If we let $\rho = 1 - \tau$ and $\sigma^2 = var(\mathbb{I})$, then for large N, the distribution of K is approximately Gaussian

$$K \sim \mathcal{N}\left(\tau N, \frac{N(\rho(1-\rho) + (\beta N)^2\sigma^2\tau\rho^2)}{(1 - \beta N \boldsymbol{E}(\mathbb{I})\rho)^2}\right). \tag{1.7}$$

Demiris (2004) explored this approximation with the multiple precision arithmetic method and validated the approximation for population sizes above 100.

Threshold modelling

As an alternative to the time-varying Poisson process for occurrence of infectious contacts, the concept of threshold to infection can also be used to define the transition from susceptible to infectious. Each susceptible individual in the population has its level of tolerance to the disease or a critical exposure to infection which represents its threshold. Therefore, an individual becomes infected at time t when the infection

pressure at time t defined by

$$A(t) = \beta \int_0^t I(s)\mathrm{d}s \tag{1.8}$$

reaches the threshold of the individual. Here again β represents the infection rate. Let Q be the random variable modelling the threshold of the susceptible individuals with Q_j be the threshold value of individual j in the population. In this model, a susceptible can be seen as an individual who has a system of defence up to a certain level and could not again fight against the disease when its defence is surpassed. In the case that the threshold distribution is exponential with parameter 1, the dynamics governing the transition from susceptible to infected are identical those described in Equation (1.2). Indeed, individual j gets infected in the small interval $(t, t+dt]$ means that he was susceptible before t. If we denote by

$$X = \Pr\left(\text{individual } j \text{ gets infected in } (t, t+dt] | j \text{ was susceptible at } t\right), \tag{1.9}$$

we have

$$\begin{aligned}
X &= \Pr\left(A(t+dt) \geq Q_j | A(t) < Q_j\right) \\
&= 1 - \Pr\left(A(t+dt) \leq Q_j | A(t) < Q_j\right) \\
&= 1 - \Pr\left(Q_j \geq A(t+dt) | Q_j > A(t)\right) \\
&= 1 - \Pr\left(Q_j \geq A(t+dt) - A(t)\right) \\
&= 1 - \exp\left\{-(A(t+dt) - A(t))\right\} \\
&= 1 - \exp\left\{-A'(t)dt + o(dt)\right\} \\
&= 1 - (1 - A'(t)dt + o(dt) \\
&= A'(t)dt + o(dt) \\
&= \beta I(t)dt + o(dt)
\end{aligned}$$

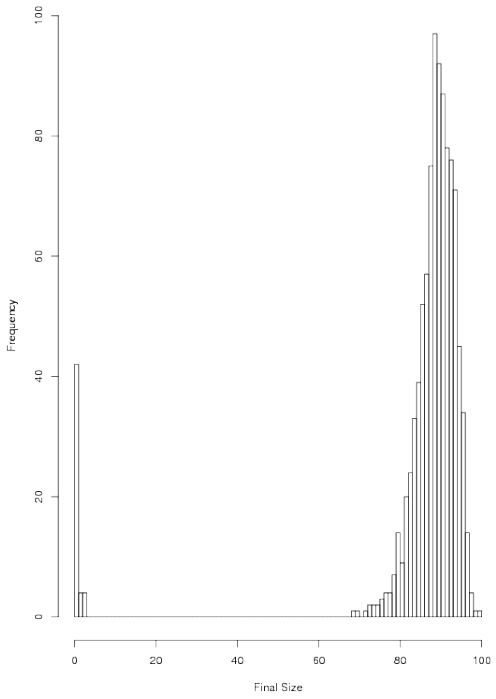
where the fourth equality uses the memoryless property of the exponential distribution, and $A'(t)$ is the derivative of $A(t)$.

We can see the equivalence between this last equation where we are interested in one particular susceptible individual and Equation (1.2). This case of Exp(1) thresholds (see Sellke, 1983) is the same as considering an exponential threshold with parameter β (the former threshold divided by β) and defining the cumulative pressure or exposure up to time t as $A_1(t) = \int_0^t I(s)\mathrm{d}s$ ($A(t)$ divided by β) as considered by O'Neill and Becker (2001). When simulating the epidemic by corresponding a threshold value to each individual, we notice that susceptible individuals become infected in an increasing order of their threshold values. We illustrate this by the following example which is considered by Streftaris and Gibson (2012) with an additional compartment of exposure time (SEIR model).

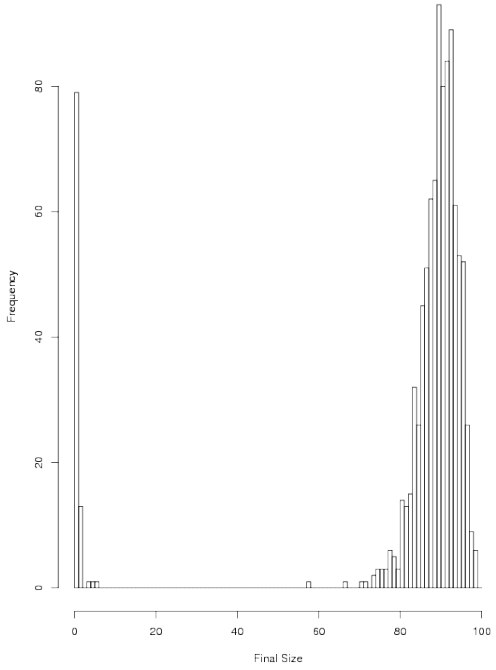
Simulation example of epidemics using threshold modelling

We set $\beta = 0.003$, $N = 100$, $a = 1$ and use a Weibull distribution for the infectious period setting it to have mean 10 and variance 100. We now choose different values for the threshold distribution parameters to study the effect of the variance on the final size. We first choose the variance of the threshold to be large, simulate 1000 final sizes and plot them in a histogram. We then choose the parameters of the threshold distribution to decrease its variance and simulate again 1000 epidemics, and repeat with decreasing variance. The histograms for four different variances of the threshold distribution while keeping the mean fixed to 1 are plotted in Figure 1.1.

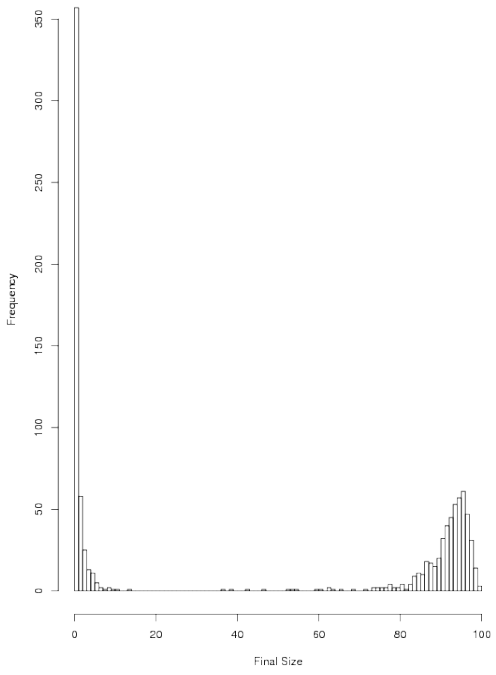
We notice that the expected epidemic size increases with the variance of the Weibull threshold distribution. The explanation is that with threshold parameters chosen such that the variance is small, and keeping the mean to be 1, all the threshold values are almost the same. Therefore, when for instance the contact parameter β is not high enough so that the cumulative pressure reaches the smallest threshold value, the epidemic does not start at all and the indexed case is quickly removed. However, when we increase β such that the epidemic starts, almost everybody becomes infected in the population. In fact, from the definition of the infection pressure, there is contribution from infected individuals in the population that increases the pressure on susceptibles. This is to say that when the number of infected individuals increases,



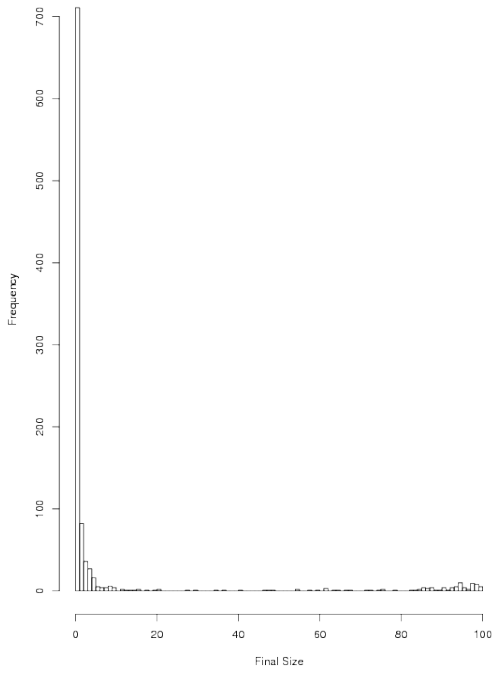
(a) $Var(Q) = 2.13$



(b) $Var(Q) = 1.58$



(c) $Var(Q) = 0.828$



(d) $Var(Q) = 0.46$

Figure 1.1: Histogram of the final size distribution using Weibull threshold with different parameters. The other parameters are $\beta = 0.003$, $N = 100$, $a = 1$ and an infectious period of mean 10 and variance 100.

the cumulative pressure increases as well. With other parameters fixed, when we modify the threshold parameters to increase the variance, the epidemic size also increases (at least on average). For large variance, the threshold values are well spread (there are small threshold values) and when the first individual gets infected, his pressure contribution helps to reach the second individual threshold and so on.

In this thesis, we will mainly use $\text{Exp}(1) \equiv \text{Weibull}(1,1)$ threshold for the model.

1.3.2 SIR model in non-homogeneous populations

We have so far described models for populations of homogeneously mixing hosts. Such an assumption, however, is often not realistic as populations regularly tend to have structure and mix heterogeneously. In this section, we briefly describe models that go beyond the homogeneous assumption by considering more realistic possibilities.

O'Neill and Becker (2001) considered a model similar to the generalised stochastic epidemic with a Gamma distribution for the infectious period but assume that the rate of contact is not constant over individuals. Indeed, a given susceptible, say j, has a tolerance to infection that is distributed according to an exponential distribution with mean $\tilde{\beta}_j^{-1}$, where $\tilde{\beta}_j$ is sampled from a $\text{Ga}(\alpha, \mu)$ distribution. This SIR model with varying susceptibility is a threshold model where each susceptible j in the population has a level of tolerance to the disease which is exponentially distributed with parameter $\tilde{\beta}_j$, where $\tilde{\beta}_j$ is sampled from a Gamma distribution.

Another model directly related to the generalised stochastic epidemic model is the multitype SIR model introduced by Hayakawa *et al.* (2003). This model considers k labelled groups of susceptible individuals. If we denote by $S_i(t)$ and $I_i(t)$, respectively, the number of susceptible and infective individuals in group i at time t $(i = 1, \ldots k)$, the total number of infectives in the population at time t is $I(t) = \sum_{i=1}^{k} I_i(t)$. The transition probabilities are defined as

$$\begin{aligned} \Pr\{S_i(t+dt) = S_i(t) - 1\} &= \beta_i S_i(t) I(t) dt + o(dt) \\ \Pr\{I(t+dt) = I(t) - 1\} &= \gamma I(t) dt + o(dt) \end{aligned}$$

where β_i is the susceptible-infective contact rate for susceptible individuals in group i. Clearly, this can be seen as a generalisation of the general stochastic epidemic where

there are different groups in the population with different known contact rate for each group. When looking at the whole population, the heterogeneity is due to this difference in the rate of contacts; however, if we concentrate on a specific group, this can be considered as a homogeneous population. Therefore, with the group number equal to 1 we are in the general stochastic epidemic case.

In the search for more realistic models, Ball *et al.* (1997) introduced the model with two levels of mixing. The model is defined in a closed population that is partitioned into groups (i.e. households or farms) of varying sizes. As in the generalised stochastic epidemic, the model allows the infectious lifetime to take any specified non-negative distribution. Individuals are allowed to mix at two levels: local and global. Contacts between individuals happen at the local and global rate according to Poisson processes. The multitype epidemic also has been extended to 2 levels of mixing by O'Neill and Demiris (2005). Very recently, the two levels of mixing model has been extended to three levels of mixing by Britton *et al.* (2011). The level added to the two levels that already exist aims to take into account secondary grouping such as school or workplace.

Other sophisticated models exist making use of network structure. Britton and O'Neill (2002) considered a population where individuals have social contacts according to a Bernoulli random graph. Albert and Barabàsi (2002) used internet networks including the so-called scale free networks.

The spread of epidemics can also be related to space. Since the seminal paper of Mollison (1977) on spatial epidemics, an increasing interest in the applied probability literature has been noticed. A number of spatial epidemic models using percolation theory (bond-percolation in particular) exist since the paper by Kuulasmaa (1982), Kuulasmaa and Zachary (1984); see also the book by Liggett (1999) and the references therein. Also, spatial models for epidemics have been studied by Grassberger (1983), de Souza and Tomé (2010), Cardy and Grassberger (1985) and Christensen *et al.* (2012).

Chapter 2

Inference tools

2.1 Introduction

This chapter describes the main background material for statistical analysis of infectious disease data. Data from infectious epidemics regularly present two main characteristics: incompleteness and inherent dependence. More often, a relatively informative dataset in epidemic modelling consists of the times at which the infectious individuals are detected. From inference viewpoint, it would be desirable to observe the times that the individuals contracted the disease, as well as the time that the individuals ended their (potential) latent period and could infect others. The incompleteness of the data occurs even more in the presence of under-reporting. The level of dependence is often related to the complexity of the model. For instance, assuming more realistic infectious periods such as Gamma or Weibull induces an additional level of dependence. Due to such features of available data in epidemic modelling, Bayesian methodology appears to be a natural framework for statistical inference. In this chapter, we first present briefly the theory of Bayesian inference. Computational methods for Bayesian inference are introduced, particularly for missing data problems.

The main parts of this chapter are organised as follows. We describe the theory of Bayesian inference in Section 2.2. The computational methodology for MCMC is discussed in Section 2.3 with the special case for missing data problems in Section 2.4. When dealing with under-reporting via MCMC, it is necessary to consider Markov chains with a state space of variable dimension, and Section 2.5 contains the techniques for this. In Section 2.6, we discuss some model reparameterisations that contribute

to a better mixing of the Markov chain. In Section 2.7, we review likelihood-based inference methods for epidemic models, particularly the Markovian SIR model. The last section is a review of statistical analysis of epidemic models; the under-reporting cases are discussed in Subsection 2.8.3.

2.2 Theory of Bayesian inference

This section describes the basic foundations of Bayesian inference. Bernardo and Smith (1994) provide a detailed and rigorous approach (see also Gelman *et al.* (2000)).

2.2.1 Bayes' theorem

In order to make inference in the Bayesian framework, there is a need for a likelihood, the conditional distribution of the data $\boldsymbol{Y}$ given the parameter(s) $\boldsymbol{\theta}$, produced by a sampling model. Unlike classical likelihood inference, Bayesian inference requires a prior distribution on the model parameters. Combining the likelihood and the prior, the posterior distribution can be derived as

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{2.1}$$

$$\propto L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.2}$$

(2.1) is referred to as Bayes' theorem. The integral in the denominator is a normalising constant and its calculation has been an obstacle until the introduction of computational techniques, namely MCMC methods (Hastings, 1970; Metropolis *et al.*, 1953; Gilks *et al.*, 1996; Gibson and Renshaw, 1998; O'Neill and Roberts, 1999). Section 2.3 will demonstrate how (2.2) can be used to obtain posterior distributions. Another interesting property of Bayes' theorem that is widely used in Bayesian inference for epidemic models particularly, is the sequential use of Bayes' theorem. In the case that

two independent data samples $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ have been collected, we have

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{Y_1},\boldsymbol{Y_2}) &\propto L(\boldsymbol{Y_1},\boldsymbol{Y_2}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= L(\boldsymbol{Y_2}|\boldsymbol{\theta})L(\boldsymbol{Y_1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto L(\boldsymbol{Y_2}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{Y_1}). \end{aligned} \tag{2.3}$$

Sequentially, the posterior distribution for the full data $(\boldsymbol{Y_1},\boldsymbol{Y_2})$ can be obtained by first evaluating $\pi(\boldsymbol{\theta}|\boldsymbol{Y_1})$, the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{Y_1}$ and then treating this posterior as a prior for the second data $\boldsymbol{Y_2}$. Equation (2.3) provides therefore a natural setting for inference when data is collected sequentially through time.

2.2.2 Prior distributions

The choice of the prior distribution is always a subject of debate. This section is focused on describing the most popular approaches for choosing a prior distribution. There exist other approaches not discussed here such as elicited priors created using expert opinions.

Conjugate priors

Computationally, some choices of prior may be more convenient than others. It is possible to select a distribution that leads to a posterior belonging to the same family as the prior. Such priors are known as *conjugate* priors. Morris (1983) showed that exponential families, to which likelihood functions often belong, do in fact have conjugate priors, so that this approach will typically be available in practice. MCMC techniques do not require the use of conjugate priors but the latter are recommended when appropriate since they provide very well known posterior distributions to sample from.

Non-informative priors

In many practical situations, prior knowledge about the parameter of interest $\boldsymbol{\theta}$ is not available. In that case there is not much to do other than specifying a non-informative prior. Also it is desirable in Bayesian inference that the information about $\boldsymbol{\theta}$ in the

posterior comes almost entirely from the data, therefore the likelihood. An example of priors that are considered as non-informative are the improper priors. These are the priors typically defined on unbounded space satisfying $\int \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \infty$. However Bayesian inference is still possible. The difficulty with improper priors is that they can lead to improper posterior distributions. This can be avoided by making sure that the normalising constant $\int L(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is finite for all $\boldsymbol{Y}$. Of course the easiest way is to use proper priors. Inferences based on improper posterior distributions are not valid.

2.2.3 Posterior distributions

Having derived the posterior distribution of interest, there are several ways in which results can be expressed. For single parameters, a plot of the posterior density is very informative and shows clearly the range of values consistent with our posterior beliefs. Indeed any summary of a distribution can be used and we briefly describe the most common used in practice.

Point estimate

Any point estimate is readily available through $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. The most commonly used are the posterior mean, the median and mode. Their use depends on the shape of the posterior distribution. For instance, the median is often preferred for one-tailed densities since the mode can be very close to non-representative values while the mean can be heavily influenced by outliers.

Interval estimation

A $100(1-\alpha)\%$ credibility interval set for $\boldsymbol{\theta}$, in the case of continuous parameter space $\boldsymbol{\Theta}$, is a subset C of $\boldsymbol{\Theta}$ which satisfies

$$1-\alpha \leq P(C|\boldsymbol{Y}) = \int_C \pi(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta}, \tag{2.4}$$

where integration is replaced by summation for discrete components of the parameter. In other words, a $(1-\alpha)\%$ credible interval is any interval whose posterior probability of containing $\boldsymbol{\theta}$ is $(1-\alpha)$. An attractive credibility set is the *highest posterior density*

set C defined by

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \pi(\boldsymbol{\theta}|\boldsymbol{Y}) \geq q(\alpha)\} \tag{2.5}$$

where $q(\alpha)$ is the largest constant satisfying $P(C|\boldsymbol{Y}) \geq 1 - \alpha$. This set can be hard to compute analytically and numerical methods are regularly used. The Bayesian credible interval appears to be a natural analogue of the frequentist confidence interval even though the concepts are different. An equal tail credibility set is mostly used by taking the $1 - \alpha/2$ or $\alpha/2$ critical points of $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$.

2.2.4 Bayesian inference for missing data problems

Missing data problems occur often in epidemic modelling particularly with under-reporting. In a Bayesian framework, these problems are typically overcome by the introduction of auxiliary variables corresponding to the missing observations.

Let $\boldsymbol{Y}_{obs}$ denote the observed data, $\boldsymbol{Y}_{mis}$ the missing data and $\boldsymbol{\theta}$ the model parameters. The distribution of $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ shares a common structure and depends on $\boldsymbol{\theta}$. In other words, the likelihood is the probability of obtaining $(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ given $\boldsymbol{\theta}$. We then have only $\boldsymbol{Y}_{obs}$ as data and $\boldsymbol{Y}_{mis}$ is treated as missing data. The pair $(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ is often referred to as the augmented or complete data. The term "missing data" can either be interpreted as data which we fail to collect for some reason or data which are not inherently available to us. In the case of under-reporting, it can be due to the inefficiency of surveillance systems. In the Bayesian framework, the following steps demonstrate how we can handle missing values problems:

$$\pi(\boldsymbol{\theta}, \boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \pi(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) \tag{2.6}$$

$$= \pi(\boldsymbol{\theta}|\boldsymbol{Y})\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}). \tag{2.7}$$

By integrating the above equation over $\boldsymbol{Y}_{mis}$, we get

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{Y})\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})\mathrm{d}\boldsymbol{Y}_{mis}. \tag{2.8}$$

We can also write

$$\pi(\boldsymbol{\theta}, \boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}), \tag{2.9}$$

and integration of the above equation with respect to $\boldsymbol{\theta}$ gives

$$\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \int \pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\mathrm{d}\boldsymbol{\theta}. \tag{2.10}$$

Equations (2.8) and (2.10) are key for sampling iteratively from the two conditional distributions of $\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs}$ and $\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs}$. This data augmentation methodology will be described in details in the MCMC algorithm in the following sections. The distribution $\pi(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$ can sometimes be referred to as posterior predictive distribution. It is used to generate multiple simulated values to replace the missing values. Samples from the joint distribution $\pi(\boldsymbol{\theta}, \boldsymbol{Y}_{mis})$ are obtained and sampling inference on $\boldsymbol{\theta}$ and $\boldsymbol{Y}_{mis}$ can be made by looking at their marginal densities. MCMC techniques are key tools that can be used.

The use of Bayesian methodology typically requires significant computational resources. In the next section, we will describe the most used computational methods nowadays, namely MCMC methods.

2.3 Markov Chain Monte Carlo Methods

2.3.1 Objective

The basic idea behind MCMC methods is to sample, at least approximately, from a specific distribution, say π, so that we can calculate functions of that distribution. π is often referred to as target distribution. In Bayesian statistics, π is the posterior distribution known up to proportionality (2.2), the exact value of the normalising constant being too difficult to calculate. There is a large number of references about the theory of MCMC and its applications (Gilks *et al.*, 1996; Brooks, 1998; Gamerman and Lopes, 2006). Regarding π as a target distribution, the key idea is to construct a Markov chain $\{X_n\}_{n\geq 0}$ with transition matrix or kernel for which π is the stationary or invariant distribution. Once the stationarity of π is ensured, then by the ergodic theorem (Norris, 1998), functions of π can be calculated. The initial state X_0 of the Markov chain does not matter. Other aspects that arise with MCMC techniques are related to the improvement of the rate of convergence of the chain if stationarity is ensured. The speed of convergence of MCMC algorithms is very important particu-

larly in real-time inference situations. MCMC methods are not restricted to finding Bayesian posterior distributions, and there are many different contexts outside the Bayesian framework where MCMC is applied. We now describe the main algorithms of MCMC methods.

2.3.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm (Metropolis *et al.*, 1953; Hastings, 1970) samples, at least approximately, a distribution π known up to proportionality without requiring knowledge of the normalising constant. The method consists of constructing a Markov chain kernel $P = (p(u, v))_{u,v}$ for the given target distribution. The constructed kernel possesses the detailed balance property as we will show below. The construction of the Markov chain is done through the specification of a proposal distribution given by its density $q(\theta, .)$. For simplicity of formulae, we will denote the distribution π as a function of the parameter only and the algorithm then follows the procedure below:

- Start from an initial arbitrary position θ_0
- To update θ_n to θ_{n+1}, generate a candidate ψ from the proposal $q(\theta_n, \psi)$ i.e choose ψ with probability $q(\theta_n, \psi)$
- Accept ψ with probability

$$\alpha(\theta_n, \psi) = \min\left\{1, \frac{\pi(\psi)q(\psi, \theta_n)}{\pi(\theta_n)q(\theta_n, \psi)}\right\} \tag{2.11}$$

  i.e set $\theta_{n+1} = \psi$ with probability $\alpha(\theta_n, \psi)$ and $\theta_{n+1} = \theta_n$ otherwise.

The transition probabilities of the resulting Markov chain are then given in general form by

$$p(u, v) = \alpha(u, v)q(u, v) = \min\left\{q(u, v), \frac{\pi(v)q(v, u)}{\pi(u)}\right\} \quad (u \neq v) \tag{2.12}$$

leading to, again for $u \neq v$,

$$\pi(u)p(u, v) = \min\left\{\pi(u)q(u, v), \pi(v)q(v, u)\right\} \tag{2.13}$$

The last equation confirms that the kernel $P = (p(u,v))_{u,v}$ possesses indeed the detailed balance equation. To obtain the correct stationary distribution, the chain is required to be irreducible and aperiodic. More details about the characteristics of the chains are available in Gamerman and Lopes (2006). The choice of the proposal distribution is not important in theory but does indeed influence the speed of convergence in practice. The most popular proposal distributions are described below.

The Independence Sampler

In this instance of the M-H algorithm, the proposal distribution is chosen to be independent of the current state. In other words, the proposal probabilities $q(u,.)$ are independent of the current state u. The consequence on the acceptance probability (2.11) is that it becomes

$$\alpha(u,v) = \min\left\{1, \frac{\pi(v)q(u)}{\pi(u)q(v)}\right\}. \tag{2.14}$$

The Metropolis Algorithm

This was the first algorithm introduced by Metropolis *et al.* (1953). The proposal distribution is such that $q(u,v) = q(v,u)\ \forall u,v$ so that the acceptance probability (2.11) becomes

$$\alpha(u,v) = \min\left\{1, \frac{\pi(v)}{\pi(u)}\right\}. \tag{2.15}$$

The proposal distribution is symmetrical and often chosen to be of the form

$$q(u,v) = g(|u-v|) \tag{2.16}$$

for some function $g(.)$ of a single variable. Such a proposal is known as symmetric random walk Metropolis (Metropolis *et al.*, 1953).

2.3.3 Gibbs Sampler

Frequently, the space on which the target distribution of interest is defined is multidimensional. Then the idea of the Gibbs sampler is to draw samples from the posterior distribution of interest using the full one-dimensional conditional distribution. More precisely if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)$ and $(\theta_1^n, \ldots, \theta_l^n)$ is the current deviate of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$

$((\theta_1^0, \ldots, \theta_l^0)$ being the initial starting values), the next update will be drawn using the full conditional one-dimensional distributions

$$\theta_i^{n+1} = \pi\left(\theta_i | \boldsymbol{Y}, \theta_1^{n+1}, \ldots, \theta_{i-1}^{n+1}, \theta_{i+1}^{n}, \ldots, \theta_l^{n}\right) \quad i = 1, \ldots, l \tag{2.17}$$

The Gibbs sampler may be viewed as a special case of the M-H algorithm, with the use of conditional distributions as alternating proposals and with acceptance probability which always turns out to be 1. For correlated parameters, convergence can be considerably improved by grouping some parameters together and update them simultaneously. The technique is known as *blocking* update (Roberts and Sahu, 1997) of parameters. In Gibbs sampling algorithm steps, it does not matter which parameters to update first.

2.3.4 Metropolis within Gibbs

This is also known as component-wise updating and is a hybrid of the Gibbs sampler and M-H. In this case the parameter space is factorised as $S = S_1 \times S_2$ where S_1 is the set of parameters for which the conditional distributions are in closed form of known distributions and S_2 is the set of parameters for which the conditional distributions are not known as coming from a particular distribution that we can straight sample from. Our aim is to use Gibbs sampler to sample from the posterior distribution π. Metropolis within Gibbs is a special case of the Gibbs sampler where a M-H algorithm can be used for sampling from the conditional distributions of those parameters that do not have a closed form to sample straight from. The Metropolis within Gibbs is used extensively in this thesis.

2.3.5 Implementation

In practice, there are some technical issues to take into account when using MCMC method to sample from a target distribution π:

- Mixing: it is important that the chain should mix reasonably rapidly. Due to high correlation between consecutive iterations, only every k^{th} sample ($k > 1$) is saved sometimes. This technique is referred to as *thinning*. Reparametrisations of the model can help to improve the mixing as we discuss in Section 2.6.

- Burn-in: the Markov chains produced by the algorithms are in fact ergodic. In other words, the distribution of $(\boldsymbol{\theta}_n)$ converges to $\pi(.|\boldsymbol{Y})$ as n tends to infinity independently of the starting value $\boldsymbol{\theta}_0$. Thus, for sufficiently large k, the resulting $(\boldsymbol{\theta}_k)$ is an approximate sample from $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. The problem in practice is to determine what "large" k means. It is advised to use diagnostic tests (Geweke, 1992; Raftery and Lewis, 1992) in the literature to assess the stationarity of the chain even though the tests do not guarantee convergence.

There exist a large number of statistical packages for implementing MCMC code and analysing output. WinBUGS (Lunn *et al.*, 2000) is the most used one and seems to be the most developed. Output of MCMC code can be analysed using statistical software such as BOA (Bayesian Output Analysis) (Smith, 2007) or CODA (Convergence Diagnosis and Output Analysis) (Plummer *et al.*, 2006).

2.4 MCMC for Missing Data Problems

2.4.1 Two-component Gibbs sampler for data augmentation

Data augmentation techniques are often used to make model likelihoods tractable. We recall that in missing data problems, the aim is to sample from the joint posterior distribution of the missing data $\boldsymbol{Y}_{mis}$ and the parameters $\boldsymbol{\theta}$. According to Equations (2.8) and (2.10), simulations from $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$ and $\pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})$ are more tractable than simulating directly from $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$. Sampling alternatively from $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$ and $\pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})$ is equivalent to Gibbs sampling methods. In particular, here we have a *two-component Gibbs sampler* which updates $\boldsymbol{Y}_{mis}$ and $\boldsymbol{\theta}$ to obtain samples from $\pi(\boldsymbol{\theta}, \boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$. In many complex models, the full conditional distributions of $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$ are often available in closed form and Gibbs sampling can be used to sample in a straightforward manner from it; but the conditional distribution $\pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})$ is not. Therefore a M-H algorithm is necessary. This goes back to the M-H within Gibbs steps described above in Subsection 2.3.4.

2.4.2 Auxiliary Variables

More generally, it is sometimes desirable to introduce an additional variable called *auxiliary*, which does not necessarily represent missing data. The approach when making inference is the same as for missing data. Let $\boldsymbol{\theta}$ represent our parameter of interest and let π_0 be the prior distribution on $\boldsymbol{\theta}$. We assume that for any values of the parameters $\boldsymbol{\theta}$, unobserved random variables $\boldsymbol{\mu}$ are generated from the density $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ and then for given $\boldsymbol{\mu}$, the observed data $\boldsymbol{Y}$ are generated in accordance with a density $h(\boldsymbol{\mu}, \boldsymbol{Y})$. Then the posterior distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{Y}$ is given by the probability function or density

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = K\pi_0(\boldsymbol{\theta}) \int_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) h(\boldsymbol{\mu}, \boldsymbol{Y}) \mathrm{d}\boldsymbol{\mu} \tag{2.18}$$

for some normalising constant K which is usually incalculable. A simple way to overcome the difficulties related to the integration is to observe that the joint posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\mu})$ is given by

$$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\mu}|\boldsymbol{Y}) = \hat{K}\pi_0(\boldsymbol{\theta}) f(\boldsymbol{\theta}, \boldsymbol{\mu}) h(\boldsymbol{\mu}, \boldsymbol{Y}) \tag{2.19}$$

where $\hat{K}$ is the appropriate normalising constant. We may then construct a Markov chain on the space of all $(\boldsymbol{\theta}, \boldsymbol{\mu})$ with the joint distribution (2.19) as its target. Ignoring the observations of $\boldsymbol{\mu}$, we can construct any desired function of θ. The method is extensively used in this thesis particularly in Chapter 4 to deal with the under-reporting problem for epidemics where missing values are obviously widely expected.

2.5 Reversible-Jump Markov Chain Monte Carlo

2.5.1 Motivation

We often need to consider Markov chains with a state space of variable dimension. This is the case with incomplete epidemics and, in this thesis, with under-reporting as the size of the epidemic is unknown.

Reversible Jump MCMC (RJMCMC) was first developed to provide a powerful tool for model selection when comparing two or more models (Green, 1995). The

parameter space for each model might have different dimensions and the idea is to obtain the posterior distributions of the parameters and the models, at the same time, so as to choose the model that provides the best fit to the data.

The RJMCMC method is an extension of the Metropolis-Hastings algorithm and its theoretical basis is well described by Green (1995). Let $\boldsymbol{Y}$ be the observed data and assume that we have a posterior distribution over parameter and model space defined up to proportionality:

$$\pi(\boldsymbol{\theta}_m, m|\boldsymbol{Y}) = L(\boldsymbol{Y}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)p(m) \tag{2.20}$$

where $L(\boldsymbol{Y}|\boldsymbol{\theta}_m, m)$ denotes the likelihood of the data given the model m and the corresponding parameter values; $p(\boldsymbol{\theta}_m|m)$ is the prior for the parameters in model m; and $p(m)$ the prior probability for model m. The RJMCMC algorithm allows us to construct a Markov chain with stationary distribution equal to the joint posterior distribution of both models and parameters, $\pi(\boldsymbol{\theta}_m, m|\ \boldsymbol{Y})$.

2.5.2 RJMCMC algorithm

The algorithm involves two steps within each iteration of the Markov chain:

- Update the parameters, $\boldsymbol{\theta}_m$, conditional on the model using the Metropolis-Hastings algorithm; and
- Update the model, m, conditional on the current parameter values. This step is where the "reversible-jump " comes in and it also consists of two steps:
    1. Propose to move to a different model with some given parameter values;
    2. Accept this proposed move with a given probability.

Let us illustrate further the steps of the reversible-jump. Suppose that at iteration i the chain is in model m with parameters $(\boldsymbol{\theta}, m)_i$ and we propose to move to model m' with parameters vector $\boldsymbol{\theta}'$. We then define a bijective function g, such that, $(\boldsymbol{\theta}', \boldsymbol{u}') = g(\boldsymbol{\theta}, \boldsymbol{u})$, where $\boldsymbol{u}$ and $\boldsymbol{u}'$ are sets of random variables with respective density function $q(\boldsymbol{u})$ and $q'(\boldsymbol{u}')$. The proposed move is accepted with probability

$$\min\left\{1, \frac{\pi(\boldsymbol{\theta}', m'|\boldsymbol{Y})P(m|m')q'(\boldsymbol{u}')}{\pi(\boldsymbol{\theta}, m|\boldsymbol{Y})P(m'|m)q(\boldsymbol{u})}\left|\frac{\partial(\boldsymbol{\theta}', \boldsymbol{u}')}{\partial(\boldsymbol{\theta}, \boldsymbol{u})}\right|\right\} \tag{2.21}$$

where $P(m|m')$ denotes the probability of proposing to move from model m to m'. The last factor in the formula above is the determinant of the Jacobian of the transformation from $(\boldsymbol{\theta}, \boldsymbol{u})$ to $(\boldsymbol{\theta}', \boldsymbol{u}')$. If the move is accepted then set $(\boldsymbol{\theta}, m)_{i+1} = (\boldsymbol{\theta}', m')$; else set $(\boldsymbol{\theta}, m)_{i+1} = (\boldsymbol{\theta}, m)_i$. In the case of model selection the posterior model probabilities can be simply estimated as the proportion of time the constructed Markov chain remains in any given model.

2.5.3 Practical considerations

In practice, several model updates may be performed within each iteration of the Markov chain. The RJMCMC algorithm possesses a lot of advantages when run for some simple cases. The acceptance probability is easy to calculate for a given proposed move in the presence of model uncertainty. Additionally, irrespective of the number of possible models within an analysis, only a single chain needs to be run to obtain the estimates of the posterior model probabilities and their corresponding parameters. With a RJMCMC algorithm we can calculate model-averaged estimates of the parameters within the Markov chain. These can simply be obtained as the estimates of the parameters, irrespective of the model for the current dimension of the chain. In summary, the reversible-jump procedure obtains estimates under individual models together with the posterior model probabilities and finally model-averaged parameter estimates, all within a single chain. However RJMCMC does not only present advantages, there are some difficulties related to it as well. It tends to essentially spend time exploring models with reasonable posterior support given the observed data, while models not supported by the data and priors are not well explored. In practice it is advisable to run a number of iterations first to eliminate models that tend to be not supported at all by the data when there is a very high number of models to select from. Another important consideration in RJMCMC is the prior specification. Care should be taken when specifying the priors on the parameters $p(\boldsymbol{\theta}|m)$ in the presence of model uncertainty since these priors can have significant impact on the corresponding posterior model probabilities by leading to opposite estimates of what should be obtained (Lindley's paradox discussed by Gilks *et al.* (1996)). Therefore prior sensitivity analysis should always be performed with different sensible priors before drawing conclusions.

2.6 Centered and Non-Centered Parameterisations

2.6.1 Parameterisations of hierarchical models

A very important feature of Bayesian modelling used throughout this thesis is non-centering parameterisations which help to improve the speed of convergence for missing data problems. Bayesian models requiring data augmentation techniques can be viewed as hierarchical models. In data augmentation, we assure that the distribution of the observed data $\boldsymbol{Y}_{obs}$ depends on the unobserved quantity $\boldsymbol{Y}_{mis}$ whose distribution depends on $\boldsymbol{\theta}$. In many contexts the *a priori* dependence between $\boldsymbol{\theta}$ and Y_{mis} can be very strong and affects the mixing of the Markov chain. The dependence $\boldsymbol{\theta} \rightarrow \boldsymbol{Y}_{mis} \rightarrow \boldsymbol{Y}_{obs}$ is called the *centered* parameterisation due to the fact that the missing data are centered between the observed data and the parameters.

On the other hand, suppose that we can find an alternative parameteristation defined by $\tilde{\boldsymbol{Y}}_{mis}$ and some function $h(.,.)$ such that $\boldsymbol{Y}_{mis} = h(\tilde{\boldsymbol{Y}}_{mis}, \boldsymbol{\theta})$ such that $\tilde{\boldsymbol{Y}}_{mis}$ is a priori independent on $\boldsymbol{\theta}$. The pair $(\tilde{\boldsymbol{Y}}_{mis}, \boldsymbol{\theta})$ is called non-centered parameterisation. There exists a mixture of both parameterisations above, called partially-non-centered parameterisation, which lies beyong the scope of this thesis. Further details can be found in Papaspiliopoulos *et al.* (2003).

2.6.2 Algorithms

The centered algorithm can be considered as the natural procedure in Bayesian missing data problems. It also provides a natural interpretation perspective for the parameters. The algorithm is a two-step procedure:

- Update $\boldsymbol{\theta}$ by drawing samples from the conditional distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs})$;
- Update $\boldsymbol{Y}_{mis}$ by drawing samples from the conditional distribution $\pi(\boldsymbol{Y}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})$.

There is also an alternative to this algorithm to improve the mixing of the Markov chain and convergence. The improvement comes from the transformation $(\boldsymbol{\theta}, \boldsymbol{Y}_{mis}) \rightarrow (\boldsymbol{\theta}, \tilde{\boldsymbol{Y}}_{mis})$ where the missing data $\tilde{\boldsymbol{Y}}_{mis}$ is a priori independent of $\boldsymbol{\theta}$. The algorithm is the following:

- Update $\boldsymbol{\theta}$ by drawing samples from the conditional distribution $\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{Y}}_{mis}, \boldsymbol{Y}_{obs})$;

- Update $\tilde{\boldsymbol{Y}}_{mis}$ by drawing samples from the conditional distribution $\pi(\tilde{\boldsymbol{Y}}_{mis}|\boldsymbol{\theta}, \boldsymbol{Y}_{obs})$.

A very good application of these algorithms can be found in Neal and Roberts (2005) where applications to epidemic models are considered. Partially non-centered parameterisations are also applied and comparisons of the speed of convergence are shown under different simulations and data applications.

2.7 Likelihood-based inference for the Markovian SIR epidemic model

In this section, we describe the likelihood derivation for the Markovian SIR epidemic and statistical inference methods based on temporal data.

2.7.1 Likelihood of the Markovian SIR model

There are three main representations adopted in the literature: the Bailey and Thomas representation (Bailey and Thomas, 1971), the Martingales-based representation (Becker, 1989) and an alternative representation proposed by Britton and O'Neill (2002) (see also Neal and Roberts, 2005). In what follows, we describe the Bailey and Thomas representation and the alternative representation. Details on the representation based on Martingales can be found in Becker (1989), Rida (1991) and Andersson and Britton (2000).

Bailey and Thomas' Representation

The transition probabilities of the model are described in Equations (1.2) and (1.3). We denote by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n_R})$, where $\tau_1 = 0$, the ordered successive removal times observed during $[0, T]$. Let ϕ_1 be the initial infection time and $\boldsymbol{\phi} = (\phi_2, \ldots, \phi_{n_I})$, the remaining successive infection times during (ϕ_1, T). Therefore, the following constraints

$$\phi_{i-1} < \phi_i < \tau_{i-2} \quad \text{for } i = 3, \ldots, n_I \tag{2.22}$$

must be satisfied, in order to obtain an epidemic.

The model is homogeneous and therefore can be fully described by the number of infected and removed individuals at each point in time, without the need of knowing

which individual got infected or removed. Also, because of the Markovian property of the model, it does not require to pair the infection and removal times. The likelihood of the model can be written as follows:

$$L(\boldsymbol{\tau}, \boldsymbol{\phi}; \beta, \gamma, \phi_1) \propto \prod_{j=1}^{n_R} \gamma R(\tau_j^-) \prod_{j=2}^{n_I} \beta S(\phi_j^-) I(\phi_j^-) \exp\left\{ - \int_{\phi_1}^{T} (\beta S(t) I(t) + R(t)) \, \mathrm{d}t \right\} \tag{2.23}$$

where the notation ϕ_j^- denotes the left limit of ϕ_j, i.e. the time just prior to ϕ_j. The likelihood in (2.23) is also given in O'Neill and Becker (2001) and O'Neill and Roberts (1999).

An alternative Representation

A rather different but equivalent representation was adopted by Britton and O'Neill (2002), Neal and Roberts (2005) and also used by Clancy and O'Neill (2008). The individuals who got infected during the epidemic are labelled as $i = 1, \ldots, n_I$ and those who did not as $i = n_I + 1, \ldots, N$. To each individual is assigned an infection (s_i) time and removal (r_i) time; assuming $s_i = \infty$, for $i = n_I + 1, \ldots N$, the uninfected individuals during the course of the epidemic. If we denote by w, the first infected individual, the likelihood of the model can be written as

$$\begin{aligned} L(\beta, \gamma; \boldsymbol{s}, \boldsymbol{r}) \quad \propto \quad & \left\{ \prod_{i=1, i \neq w}^{n_I} \beta I(s_i^-) \right\} \exp\left(-\beta \int_{s_w}^{T} S(t) I(t) dt \right) \\ & \gamma^{n_R} \exp\left(-\gamma \sum_{i=1}^{n_R} (r_i - s_i) \right) \exp\left(-\gamma \sum_{i=n_R+1}^{n_I} (T - s_i) \right). \end{aligned} \tag{2.24}$$

It can be shown (see Neal and Roberts, 2005) that

$$\int_{s_w}^{T} S(t) I(t) \mathrm{d}t = \sum_{i=1}^{n_I} \sum_{j=1}^{N} (r_i \wedge s_j - s_j \wedge s_i) \tag{2.25}$$

where the notation $\wedge$ denotes the minimum.

From a practical point of view, Equation (2.25) presents the advantage of avoiding discretisation to transform the integral into sum, such discretisation being substituted by the double sum.

2.7.2 Inference for the Markovian SIR epidemic

Case of complete data

For complete data, MLEs for the parameters can be derived by differentiating the likelihood. However in this thesis we are focused on Bayesian inference.

The Bayesian approach can be adopted by assigning (conjugate) Gamma prior distributions for β and γ:

$$\beta \sim \text{ Ga}(\nu_\beta, \lambda_\beta),\ \gamma \sim \text{ Ga}(\nu_\gamma, \lambda_\gamma). \tag{2.26}$$

We apply Bayes theorem by multiplying the prior distributions and the likelihood (using Bailey and Thomas representation) and obtain the joint posterior distribution of β and γ:

$$\begin{aligned} \pi(\beta, \gamma|\boldsymbol{\tau}, \boldsymbol{\phi}) &\propto \beta^{\nu_\beta + n_I - 2} \exp\left\{-\beta\left(\int_{\phi_1}^{T} S(t)I(t)\mathrm{d}t + \lambda_\beta\right)\right\} \\ &\times \gamma^{\nu_\gamma + n_R - 1} \exp\left\{-\gamma\left(\int_{\phi_1}^{T} I(t)\mathrm{d}t + \lambda_\gamma\right)\right\}. \end{aligned} \tag{2.27}$$

The two parameters are *a posteriori* conditionally independent and their conditional distributions are given by

$$\beta|\boldsymbol{\phi}, \boldsymbol{\tau} \sim \text{Ga}\left(\nu_\beta + n_I - 1, \lambda_\beta + \int_{\phi_1}^{T} S(t)I(t)\mathrm{d}t\right), \tag{2.28}$$

$$\gamma|\boldsymbol{\phi}, \boldsymbol{\tau} \sim \text{Ga}\left(\nu_\gamma + n_R, \lambda_\gamma + \int_{\phi_1}^{T} I(t)\mathrm{d}t\right). \tag{2.29}$$

It is therefore straightforward to sample from the posterior distribution of the two parameters β and γ and any function of the two parameters, such as the basic reproduction number, R_0.

Case of partially observed epidemics

We assume that only the removal times of the infected cases are observed and treat the infection times as model parameters. Here we adopt the alternative representation and again consider Gamma prior distributions for β and γ. The full conditional

posterior distributions are given by

$$\beta|\boldsymbol{r},\boldsymbol{s},\gamma \sim \text{Ga}\left(\nu_\beta + n_I - 1, \lambda_\beta + \int_{s_w}^{T} S(t)I(t)\text{d}t\right), \tag{2.30}$$

$$\gamma|\boldsymbol{r},\boldsymbol{s},\beta \sim \text{Ga}\left(\nu_\gamma + n_R, \lambda_\gamma + \sum_{i=1}^{n_R}(r_i - s_i) + \sum_{i=n_R+1}^{n_I}(T - s_i)\right). \tag{2.31}$$

The rates β and γ can then be updated with Gibbs sampling steps. It remains to estimate the infection times and we do this using a Metropolis-Hastings algorithm. The methodology here is an example of Metropolis-within-Gibbs algorithm described in Subsection 2.3.4. The infection times are updated according to the state of the individuals in the population and there are 3 of them denoted as follows:

- 0 - Susceptible
- 1 - Removed
- 2 - Infected but not removed before T.

Following closely the algorithm described by Streftaris and Gibson (2004a), the infection times can be updated as follows:

- Choose one individual in the population at random (let us say k)
- If the state of the individual is 1 which means the individual was infected and removed before time T we simply update its infection time uniformly in (T_0, r_k), where T_0 is the lower bound for the infection times and r_k is the observed removal time for individual k. The proposed infection time is accepted with probability

$$A_{1\to 1} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})}\right\} \tag{2.32}$$

  We can propose the new infection time s_k such that $(s_k - r_k) \sim \text{Exp}(\gamma)$. Therefore the acceptance probability becomes:

$$A'_{1\to 1} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})} \times \frac{\exp\{-\gamma(r_k - s'_k)\}}{\exp\{-\gamma(r_k - s_k)\}}\right\} \tag{2.33}$$

  where s'_k is the current infection time of individual k.

- If the state of the individual is 0 which means the individual is still susceptible, we propose to add an infection time s_k uniformly chosen in (T_0, T). On the reverse move we propose with probability 0.5 to delete the added infection time. Therefore the acceptance probability is

$$A_{0\to 2} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})} \times \frac{T-T_0}{2}\right\}. \tag{2.34}$$

  If the infection time is added, the state of the individual becomes 2.

- If the state of the individual is 2, with probability 0.5 we propose either to delete the infection time added or to update it. We delete the added infection time with probability

$$A_{2\to 0} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})} \times \frac{2}{T-T_0}\right\}, \tag{2.35}$$

  and the state of the individual becomes 0.

  The added infection time is updated by proposing a new time uniformly chosen in (T_0, T), and accept it with probability

$$A_{2\to 2} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})}\right\}. \tag{2.36}$$

If the epidemic is known to have ceased, we are only interested in the state 1 individuals and the algorithm is therefore considerably reduced to one step, i.e. updating the infection times of the removed individuals.

2.8 Review of Statistical analysis in epidemic models

Statistical analysis plays an important role in bridging the gap between the mathematical theory of epidemics and public health. In this section, we review the methods of parametric inference on model parameters and epidemiologically important parameters. Becker (1989) provides a detailed inference method for non-parametric methodology using Martingales.

2.8.1 Inference on epidemics in homogeneous populations

Statistical inference uses the likelihood function and in the frequentist approach, by differentiating the likelihood, it is straightforward to obtain maximum likelihood estimates (MLE) for the parameters of interest. Bailey and Thomas (1971) employed this methodology to estimate infection and removal rates on a continuous time model. Rida (1991) derived asymptotic normality and consistency for the MLEs of the infection and removal rates with the corresponding reproduction number. The standard errors were derived and approximate confidence intervals can be obtained by normal approximation. Becker (1989) provides the largest amount of information for inference based on epidemic models in homogeneous populations. The nature of data in epidemic modelling makes statistical inference not straightforward to perform in the classical approach. When the epidemic is partially observed, then the likelihood cannot be written in closed form. Becker (1989) provides an estimate for the infection rate using martingale theory.

The Bayesian framework offers a natural methodology for missing data problems (see Subsection 2.2.4). Inference for stochastic epidemic models, as many applications in statistics, has benefited from the use of MCMC methods. The first statistical analysis of SIR models using MCMC methods were presented by Gibson and Renshaw (1998) and O'Neill and Roberts (1999). Departing from the Markovian SIR model, different distributions were assumed for the infectious lifetime period, notably Gamma distribution (see O'Neill and Becker, 2001; Jewell *et al.*, 2008; Clancy and O'Neill, 2008) and Weibull distribution (Streftaris and Gibson, 2004a). An extension of the the basic homogeneous model is made by Hayakawa *et al.* (2003) to allow multitype (e.g. sex, age) model, with a different rate of infection between types, and an unknown actual number of susceptible individuals. Statistical inference is therefore required for the infection rates and the population size.

2.8.2 Epidemics in structured population

The assumption of homogeneous mixing populations is not realistic in many applications. We briefly discuss here some references where statistical analysis were performed in populations with different structures. Many models in the literature allowing heterogeneity between the hosts are disease-specific and usually take into

account many other compartments and put accent on the covariates specific to the disease. Another approach when studying heterogeneity is to divide the population into groups where individuals mix homogeneously within each group. A *Mixing matrix*, whose elements r_{ij} specify the probability that an individual in group i will have a potentially infectious contact, is used to model the contacts between groups. More complex structure of the mixing matrix have been considered by Koopman *et al.* (1989).

Structured populations have been analysed statistically in many ways. Longini and Koopman (1982) studied models in which individuals reside in households and may potentially be infected from infectives within their household or different households. They assume that the disease within the household progresses independently of the dynamics of the community. One model of this kind is the extension by Addy *et al.* (1991) of the work in Ball (1986) on the generalised stochastic epidemic so that individuals can also be infected from the community at large. Britton and Becker (2000) use the model of Longini and Koopman (1982) to estimate the critical vaccination coverage required to prevent epidemics in a population partitioned into households. O'Neill *et al.* (2000) applied MCMC methods to analyse temporal and final size data. Other work involving vaccination strategy includes Ball and Lyne (2002) who derived the effect of different vaccination policies in a population that is partitioned into households. Becker *et al.* (2003) use an independent households model to estimate vaccine efficiency from household outbreak data.

Ball *et al.* (1997) introduced epidemics with two levels of mixing and their statistical inference. The model assumes local and global contacts and the statistical inference discusses various vaccination strategies. Bayesian inference for this two level of mixing model is available in Demiris and O'Neill (2005). An extension to the two levels of mixing is considered by Britton *et al.* (2011) where a further level of mixing is considered. Such further level models secondary grouping like school or workplace while household and the whole community represents the other two levels. An MCMC method on final size data is applied after derivation of the likelihood.

Despite all these efforts to model real life applications, it is easy to realise the practical impossibility to capture every aspect of the structure of populations. Therefore researchers have been concerned with modelling the population structure through ran-

dom networks. Britton and O'Neill (2002) use MCMC methods to conduct a Bayesian inference for a model where individuals have social contacts according to a Bernoulli random graph. Extension to more complicated social structures and networks are of interest as the review by Albert and Barabàsi (2002) pointed out. Also, there has been some statistical inference for spatio-temporal epidemic models, see for example Gibson (1997); Marion *et al.* (2003); Filipe *et al.* (2009).

2.8.3 Applications in the case of under-reporting

References in the literature treating under-reporting situations in epidemics are not frequent. During the pandemic of the avian influenza A (H1N1), Fraser *et al.* (2009) estimated the reproduction number from the time series of cases given. Their general approach is to develop a statistical framework to estimate the time-dependent reproduction number R_t in the context of variable reporting rates. Working in a Bayesian framework, Fraser et al. assumed a fixed serial interval distribution (distribution of time periods between subsequent infections in a chain of transmissions) and accounted for under-reporting by explicitly modelling the reporting process. By denoting by N_t the number of detected cases with symptoms onset at time t, M_t the unobserved "real" number of cases with symptoms onset at time t and ρ_t the reporting rate at time t, Fraser et al. assumed a Binomial distribution for the detected cases

$$N_t \sim \text{Bin}(M_t, \rho_t). \tag{2.37}$$

This is well detailed in the supporting material of the early findings paper Fraser *et al.* (2009). A similar approach will be considered in this thesis based on temporal data. Hens *et al.* (2011) also applied a fixed serial interval distribution and accounted for under-reporting while casting it in the likelihood (frequentist) framework. Also, in the frequentist framework and using daily cases of the epidemic, White and Pagano (2010) considered a likelihood-based methodology to investigate the impact of under-reporting on estimates of both R_0 and the serial interval. Recently, work by Dorigatti *et al.* (2012) couples a deterministic mathematical model with a statistical description of the reporting process, with application to a surveillance data collected in Italy, again for H1N1 influenza epidemic. The reporting rate was assumed to be age-

dependent and estimation was performed via MCMC method. Before all the above references, Clarkson and Fine (1985) examined methods for estimating the efficiency of measles and pertussis notification (reporting) in England and Wales. Using time series data, estimates were obtained from a comparison of annual number of births and notifications with modification of the approach to include detailed age-specific data. Their estimated reporting rate (just over 50% for measles) was used to correct the under-reporting through a process of susceptible reconstruction (see Bjornstad *et al.*, 2002).

2.9 Conclusion

The models described in Section 1.3 and their statistical analysis described in the previous section play a significant role in the theory of modelling and contribute to policy decisions. This thesis does not attempt to make further studies on these models. We consider mainly one model, the general stochastic model, and incorporate under-reporting in order to study the effect of ignoring existence of under-reporting. Problems of under-reporting are not widely studied in the literature. We go further to provide statistical methods in the Bayesian framework, treating the non-reported events as missing data (Chapter 3). We also consider the use of approximations which help to speed up the inference process as we will see in Chapter 4.

Chapter 3

Epidemics with constant probability of reporting

3.1 Introduction

In this chapter, we explore the effect on statistical inference of epidemics with under-reporting. The aim of the study is twofold: Firstly we investigate the possibilities of bias in estimation if we treat the data as if no under-reporting were occurring. We expect that this will, in general, lead to underestimation of infection rates and reproduction number. Secondly, assuming we make allowance for the fact that under-reporting is occurring, we develop natural approaches to show how and how well we can estimate the rate of under-reporting and other epidemiological parameters in the model.

For example in the recent influenza A (H1N1) flu pandemic, the question of under-reporting was crucial. Early findings based on methodology developed by Fraser *et al.* (2009) considered the daily cases observed with potential under-reporting rate to estimate the reproduction number at each given day. Their approach is mainly based on the observed final size which is binomially distributed, conditional on the actual final size, with parameters given by the true unobserved number of cases and the rate of reporting. Under-reporting was also considered by Hens *et al.* (2011) where a non-parametric approach was used. Also by looking at the daily number of cases, they presmoothed the cumulative number of cases based on a non-parametric model to infer the missing update.

Our approach in this paper is based on Bayesian methodology using temporal data. The main focus of this study is the bias that under-reporting may introduce in the estimation of model parameters in general and particularly in estimates of the reproduction number. We are thus driven to consider how one can overcome potential bias by incorporating the reporting process in the model and adjusting MCMC updates accordingly.

In the remainder of this chapter, we will first describe a general framework for modelling epidemics emphasising threshold models. The focus will then be on the Markovian model, i.e. the general stochastic epidemic with a reporting process incorporated in. It is important to point out that the physical progression and the reporting process in the model are independent meaning that the reporting process does not affect the dynamic evolution of the disease and vice versa. Section 3.4 will contain the RJMCMC algorithm used to make inference. Applications to data will be presented in Section 3.5. After the description of the data, we will show how much can be lost in the quantification of R_0 if under-reporting is ignored. Results using algorithm from Section 3.4 will be presented with conclusions in Section 3.6.

3.2 General framework for modelling an SIR epidemic

Epidemics have been widely modelled by dividing populations into compartments. One of these compartmental models frequently used is the SIR model. The idea here is to develop a general framework to characterise in a unique way the SIR model. The models described in this framework are not new, but the description below contributes to a better understanding of how epidemics can be modelled. An easy extension can be made to models with more compartments.

3.2.1 Modelling the SIR epidemic

Let us assume that there are N individuals in the population where each potentially undergoes transitions $S \to I \to R$.

The *history* $\mathcal{F}_t$ of the process at any time $t \geq 0$ is given by the record, for each individual, of its transition times ($S \to I$ and $I \to R$) prior to time t (where these exist).

For times t_1 and t_2 such that $t_1 < t_2$, we have the relation $\mathcal{F}_{t_1} \subset \mathcal{F}_{t_2}$. In probability theory, this can be defined rigorously (see Øksendal, 2003) as the information filtration, but we can simply understand here that $\mathcal{F}_t$ is the possible information from the process through all the time period until t. As usual let $S(t)$, $I(t)$, and $R(t)$ denote the number of individuals in each state at time t.

A broad class of models for each individual i, $i = 1, \ldots N$, and each time $t \geq 0$ is given by the following. If in state S, an individual i makes transitions to state I at rate $\beta_i(\mathcal{F}_t)\eta_i(\mathcal{F}_t)\xi_i(\mathcal{F}_t)$. The term $\beta_i(\mathcal{F}_t)\eta_i(\mathcal{F}_t)$ may be regarded as the *susceptibility* of individual i at time t where $\beta_i(\mathcal{F}_t)$ is the parameter modelling all the contacts established by individual i until time t and $\eta_i(\mathcal{F}_t)$ may be regarded as any susceptibility factor of individual i at time t. The contacts established by an individual can be thought of part of its susceptibility in general but many other factors can influence the susceptibility for instance biological factors of each individual. Such factors are specific to each individual. The term $\xi_i(\mathcal{F}_t)$ represents the total *infectivity* in the system at time t acting on i. Typically $\xi_i(\mathcal{F}_t) = \xi(.)I(t)$ for some function $\xi(.)$ when the infectivity is constant with respect to infectious individuals. Note that $\xi(.)$ can be a function depending for example on factors affecting infectivity (as in Streftaris and Gibson, 2004b). An individual i, if in state I, makes transitions to state R at rate $\gamma_i(\mathcal{F}_t)$, where $\gamma_i(.)$ is typically the *hazard rate function* associated with the infectious lifetime distribution of individual i and is independent of everything else. We assume that there is no other factor influencing the transitions apart from the ones discussed.

We describe models as *homogeneous* if individuals are essentially indistinguishable, i.e. the functions β_i, $\eta_i(.)$ and $\gamma_i(.)$ are the same for each i, depending only on the history of i and exchangeably of the set of histories for the remaining individuals.

3.2.2 Threshold Model

A special case of the general framework above is the threshold modelling. Sellke (1983) was the first to formulate a SIR compartmental model using the threshold concept for the infectious process. Each susceptible individual in the population has its level of tolerance to the disease or a critical exposure to infection which represents its threshold. Therefore, an individual becomes infected at time t when the total

infection pressure exerted on it at time t reaches its threshold.

A fairly general homogeneous model, in this framework, is given by taking $\beta_i = \beta$ and

$$\eta_i(\mathcal{F}_t) = \rho\left(\beta \int_0^t I(u)\, du\right) \tag{3.1}$$

$$\xi_i(\mathcal{F}_t) = \sum_{j \in I(t)} \xi(t - t_j) \tag{3.2}$$

$$\gamma_i(\mathcal{F}_t) = \gamma(t - t_i), \tag{3.3}$$

where $\rho(.)$ is the hazard rate function of the threshold distribution and $\gamma(.)$ is the hazard rate function of the infectious lifetime distribution, $\xi(.)$ is a general function. β is the contact rate so that

$$A(t) = \beta \int_0^t I(u)\, du \tag{3.4}$$

is the infection pressure exerted until time t on a given individual, and t_i is the point of time at which individual i is infected. This model takes into account the possibility of varying infectivity among the different infectives in the population. The threshold of each individual, denoted by Q_i for individual i, is an unknown quantity and is therefore modelled as coming from some probability distribution with hazard rate function ρ.

In Sellke's construction, the individual threshold levels Q_i are assumed to be Exp(1) random variables, facilitating the model construction from a mathematical point of view. But in this general framework here, we allow the choice of any positive random variable for the tolerance level. The dynamic transition from one compartment to another can explicitly be written, thus the likelihood can be derived.

3.2.3 Example

As an example, let us assume that the threshold of each individual comes from a Weibull distribution (Streftaris and Gibson, 2012). The density of the Weibull distri-

bution parameterised as

$$f(x, \nu, \lambda) = \nu\lambda x^{\nu-1} \exp(-\lambda x^{\nu}) \text{ for } x \geq 0 \text{ and } \lambda, \nu > 0, \tag{3.5}$$

gives a hazard rate function

$$\rho(x) = \nu\lambda x^{\nu-1}. \tag{3.6}$$

Therefore, if we assume that the function ξ is constant and $\xi = 1$, the rates $\eta_i(\mathcal{F}_t)$ and $\xi_i(\mathcal{F}_t)$ become

$$\eta_i(\mathcal{F}_t) = \nu\lambda \left(\beta \int_0^t I(u)\, du\right)^{\nu-1} \text{ and } \xi_i(\mathcal{F}_t) = I(t). \tag{3.7}$$

Thus, with $\beta_i(\mathcal{F}_t) = \beta$, the transition probability of the infection process is

$$Pr\,(j \text{ gets infected in } (t, t+dt) | j \text{ was susceptible at } t) =$$
$$\nu\lambda \left(\beta \int_0^t I(u)\, du\right)^{\nu-1} \beta I(t)\, dt + o(\, dt). \tag{3.8}$$

The example here is also applicable to the case where a Weibull distribution assumption is made for the infectious lifetime for the disease. Then the probability dynamic of transition $I \to R$ can be written as

$$Pr\,(j \text{ gets removed in } (t, t+dt) | j \text{ became infected at } t_j) = \nu\lambda(t - t_j)^{\nu-1}\, dt + o(\, dt). \tag{3.9}$$

We now suppose in the model defined by (3.1)–(3.3) that the function ρ is constant and equal to unity. Then we may think of the model as being given by each individual i, while infectious, infecting each remaining individual at rate $\xi(t - t_i)$ and being removed at rate $\gamma(t - t_i)$. If we are interested only in the distribution of the number of individuals eventually infected, then the model reduces to a (homogeneous) graph model, in which every site (individual) is a neighbour of every other. In the case the graph link between vertices is random, we therefore have an epidemic on a random graph model. This connection to the epidemic on graph models needs further discussion which will lead to social networks and spatial models.

Clearly, the modelling here allows for a lot of flexibility in the choice of the dis-

tributions for the threshold and the infectious period. It is then obvious to raise the question of distribution choice. This issue is not considered in this thesis, but it is further discussed by Streftaris and Gibson (2012).

The transmission of individuals from one state to another that we use in this chapter fits well in the framework described above. It is actually one of the simplest approaches as we describe below.

3.3 Markovian SIR model and reporting process

This section considers a specific case of the framework described above for the physical progression with the reporting process added in the modelling.

For the purpose of making inference, the model needs to be regarded as having two components: the transition of individuals from one state to another, which we call physical progression of the epidemic, and the case-reporting or observation process. We assume throughout this work that the disease transmission is not influenced by the reporting process. In other words, the spread of the epidemic itself is not influenced by the reporting. This assumption is more realistic if the reporting in the model happens at removal times. In many situations, the spread of the epidemic is supposed to be dependent on the reporting process. For instance in large populations, when the reported cases reach a certain level, change of behaviour in the population would contribute to a modification of the process for the physical epidemic. Also, different campaign of information about the epidemic would affect the behaviour and therefore the underlying process of the spread of the disease. For small population sizes and when the epidemic is fast and there are no surveillance techniques and no policy measure against the disease, the physical progression of the epidemic would not be affected by the reporting. On the other hand, questions of change of behaviour would be more frequent to happen in the case the reporting happens at infection times. In this study, the emphasis is on the effect of under-reporting and therefore probable changes of behaviour are not considered.

3.3.1 Markovian SIR epidemic

The Markovian SIR epidemic is also known in the literature as the general stochastic epidemic and we first describe it in the framework in Section 3.2.

The susceptibility of an individual can be thought of as the rate β of contact per susceptible-infectious in the population O'Neill and Becker (2001). If we assume constant infectivity from all infectious individuals, the total infectivity acting on an individual at time t can be regarded as the number of infectives in the population at that time. We obtain that $\beta\eta_i(\mathcal{F}_t) = \beta$, $\xi_i(\mathcal{F}_t) = I(t)$ and the probability dynamic for $S \to I$ is

$$\Pr(j \text{ gets infected in } (t, t+dt) | j \text{ was susceptible at } t) = \beta I(t)\, dt + o(\, dt). \quad (3.10)$$

This transition dynamic is equivalent to considering an Exp(1) threshold in Sellke's construction i.e $\rho = 1$ $\beta_i = \beta$ and $\xi_i = 1$ in the threshold modelling in Subsection 3.2.2 with Equations (3.1) and (3.2).

When we consider the rate $\gamma_i(\mathcal{F}_t)$ as constant, say γ, for each infected individual in the population, we get the transition probability for $I \to R$

$$\Pr(j \text{ gets removed in } (t, t+dt) | j \text{ is still infectious at } t) = \gamma\, dt + o(\, dt) \quad (3.11)$$

Equation (3.11) is equivalent to considering Exp(γ) infectious periods. The memoryless property of the exponential distribution makes the model Markovian. The work produced here can be extended to other distributions for the infectious lifetime.

The formulations of the two processes of infection (3.10) and removal (3.11) are individually-based models i.e each individual in the population is labelled and processes are defined with respect to them. Considering the whole population and defining the same processes for the Markovian SIR epidemic is equivalent to formulating the transition dynamics as

$$\Pr(S(t+dt) = S(t) - 1) = \beta S(t) I(t)\, dt + o(\, dt) \quad (3.12)$$

$$\Pr(I(t+dt) = I(t) - 1) = \gamma I(t)\, dt + o(\, dt). \quad (3.13)$$

Combining with the Markovian SIR model, we now consider and model the reporting process.

3.3.2 The reporting process

There are many possibilities for the nature of the reporting process as we describe below.

The base model

In a Markovian SIR model, we assume that all infection times are unknown. We further assume that each removal event is independently reported with probability p. We point out that the reporting probability p is constant and is therefore independent of both individuals and time. This first possibility means that for infected individuals, their removal times are reported with probability p, but their infection times are not. This implies that with probability $1-p$, no event regarding these individuals is reported, and it is not even known whether or not they have become infected. Since only removal times are observed, we are assuming that some hidden removals have occurred and that the reporting is not affecting the course of the epidemic's spread.

The assumption of constant probability of reporting leads us to derive the following distribution for the observed number of reported cases which we denote by n_{rep}. Let n be the unknown number of removals with full case or perfect reporting. Then, due to the independence of the reporting between individuals, n_{rep} conditional on n, is binomially distributed with parameters n and p

$$n_{rep}|n \sim \text{Bin}(n, p). \tag{3.14}$$

For inference purpose, this means that the number of reported removed individuals is a binomial proportion p of the true number of removed individuals.

On the other hand, from equation (3.13), the removal process follows a non-homogeneous Poisson process with intensity $\gamma I(t)$. Making use of the random selection property of a Poisson process Kingman (1993), the new process obtained for the reported removal times is Poisson with intensity $p\gamma I(t)$. The main difficulty about using this new Poisson process to obtain the likelihood function is the fact that $I(t)$

is unknown.

Variants of the model

In reality, the infection times during epidemics are usually not observed. It is therefore realistic to assume that only removal times for reported individuals are known. The model here aims to study the effect of under-reporting with having the physical progression of the epidemic unchanged. Therefore we consider different possibilities by being flexible in reporting assumptions. The constant probability of reporting assumes that there is no event influencing the reporting process throughout the course of the epidemic. In later sections we will consider factors affecting the reporting process, resulting in more realistic approaches (for some circumstances).

We may also assume that complete information is available for reported individuals, namely both their infection and removal times. This assumption means that with probability p we know the infection and removal time of each infected individual and with probability $1-p$, neither of these event times is known. The difficulty with this assumption is that in reality such individuals whose infections are observed would be subject to treatment.

One of the very important issues with epidemic modelling is the information we can obtain as time evolves. How much we can learn as the epidemic evolves through time is a crucial question if we want to apply the model to real cases of ongoing epidemics. It is then important to have a clear idea of how these epidemics evolve and to develop methods that take into account the time observation period of the epidemic for ongoing diseases.

We can combine all the different variants with the base model and thus consider the following 4 main inferential problems:

- with a completed epidemic, only a binomial proportion of the removal times has been reported and in this case, all the infection times are unobserved;
- only a binomial proportion of removal times has been reported but the epidemic is incomplete;
- a binomial proportion of the pair of infection and removal times are observed: either the reporting happens at infection times and each infection time is ob-

served with probability p while the removal time is observed with probability 1 or the removal is observed with probability p and the infection time of the reported individual is assumed to be known for some reason (example of contact tracing (Eames and Keeling, 2003; House and Keeling, 2010)). This case is mainly treated in chapter 4;

- a binomial proportion of the pair infection-removal times has been reported but the epidemic is incomplete.

In many of the cases described here, where the physical progression of the epidemic is unchanged, the likelihood function using data augmentation is the same. The differences come from the variables used to augment the data, which are not the same for all the models considered as described in the the following subsection.

3.3.3 Likelihood function

Due to the unobserved times we need to augment the data to make the likelihood function tractable. We denote by $\boldsymbol{r} = (\boldsymbol{r_o}, \boldsymbol{r_u})$ the vector of removal times at the end of the observation period T where $\boldsymbol{r_o}$ and $\boldsymbol{r_u}$ are the vectors of observed and unobserved removal times respectively. Similarly, we denote by $\boldsymbol{s} = (\boldsymbol{s_o}, \boldsymbol{s_u})$ the corresponding infection times. Let $\mathcal{I}$ and $\mathcal{R}$ be respectively the sets of all infections and all removals that happen before T, and $\bar{\mathcal{R}}$ be the set of individuals infected but not removed before T. We denote by w the first infected individual in the population; the set $\mathcal{I}_{-w}$ denotes all the infected individuals excluding w, and $\boldsymbol{s}_{-w}$ the vector of infection times without the first infection. By using n_{rep} to denote the number of reported removals, the likelihood function can be written as

$$\begin{aligned} L(\beta, \gamma, p; \boldsymbol{s}_{-w}, s_w, \boldsymbol{r}) \;\propto\; & \left\{ \prod_{i \in \mathcal{I}_{-w}} \beta I(s_i^-) \right\} \exp\left(- \int_0^T \beta S(t) I(t) dt \right) \\ & \prod_{i \in \mathcal{R}} \gamma \exp\left(-\gamma(r_i - s_i)\right) \prod_{i \in \mathcal{I} \cap \bar{\mathcal{R}}} \exp\left(-\gamma(T - s_i)\right) \quad (3.15) \\ & p^{n_{rep}} (1-p)^{|\mathcal{R}| - n_{rep}} \end{aligned}$$

where $|\mathcal{R}|$ is the unknown true number of removals that happened before T and s_i^- denotes the left limit of s_i, i.e. the time just prior to s_i.

The first two lines of the likelihood function in 3.15 carry information from the Markovian SIR formulation while the last line reflects the reporting process. Due to the independence between the two events of physical progression and reporting, the likelihood is simply the product of the probabilities of occurrence of each of them. If we assume that the infection times are known for the reported cases, the augmentation of the data will only come from the inclusion of event times (removal and infection) of non-reported cases.

3.4 Inference

The aim is to study inference possibilities under the 4 inferential problems we discussed earlier. It is important to recall that the derived likelihood involves unobserved events and therefore estimation must involve data augmentation techniques. We adopt a Bayesian methodology as it provides a natural framework to incorporate prior knowledge and treat quantities that are not observed as parameters in the model. We then need to specify priors on our parameters and derive the posterior distributions. The sampling from the posterior distributions are done using MCMC methods, more precisely Metropolis-Hastings with Gibbs updates. The other aspect of the updates is the unknown number of total infections which require us to perform RJMCMC algorithm instead and we provide more details below.

3.4.1 Updates of parameters β, γ

The gamma distribution is a conjugate prior for β and γ in likelihood (3.15). Assuming

$$\beta \sim \text{Ga}(\nu_\beta, \lambda_\beta), \text{ and } \gamma \sim \text{Ga}(\nu_\gamma, \lambda_\gamma) \tag{3.16}$$

lead to the following full conditional posterior distributions:

$$\beta | \boldsymbol{r}, \boldsymbol{s}_{-w}, s_w, \gamma \sim \text{Ga}\left(\nu_\beta + |\mathcal{I}| - 1, \lambda_\beta + \int_0^T S(t)I(t)dt\right), \tag{3.17}$$

$$\gamma | \boldsymbol{r}, \boldsymbol{s}_{-w}, s_w, \beta \sim \text{Ga}\left(\nu_\gamma + |\mathcal{R}|, \lambda_\gamma + \sum_{i \in \mathcal{R}} (r_i - s_i) + \sum_{i \in \mathcal{I} \cap \bar{\mathcal{R}}} (T - s_i)\right). \tag{3.18}$$

Therefore β and γ can be updated through a Gibbs sampling steps.

The updating of the infection and removal times requires a change of dimensions due to the unknown number of infections and removals. We implement a reversible jump Markov chain Monte Carlo algorithm (Section 2.5) where we need to add, remove or move the removal and infection times that are not observed. The algorithm given in the following section for the time updates can be viewed as a generalisation of the algorithm described by Streftaris and Gibson (2004a).

3.4.2 Reversible Jump MCMC algorithm for reporting process

The algorithms for updating the unobserved event times are described below. Firstly the description only takes into account complete epidemics meaning that the epidemic is known to have ceased. We then move on to extend it to incomplete or ongoing epidemics. The case of incomplete epidemic algorithm coincides with the case of complete epidemic when $T \to \infty$. The algorithms are based on the states characterising each individual. It is natural to question how to know if an epidemic is complete when we have under-reporting. For an epidemic that happened in the past a long time before inference including under-reporting, it is easy to assume a complete epidemic. In other cases where the epidemic is recent, the rate at which reporting occurs can help to identify if the epidemic is completed or not. For instance if we do not have any reported case for a long period of time, we can assume that the epidemic has finished. How the long period of time is determined is another issue and can be related to the disease in question. Of course a better conclusion will be surely made with a surveillance system.

RJMCMC for the case of complete epidemic

The RJMCMC algorithms in this work are focused on the base model where reporting happens at the removal times. However the algorithms below are well applicable to other variants of the base model by simply renaming the states since the basic idea about the updates and the acceptance probabilities are the same. The case of complete epidemic requires 3 states $\{0, 1, 2\}$, providing details on individuals as follow:

- 0 - Susceptible

- 1 - Removed and reported
- 2 - Infected and removed but not reported

The states above condition the moves in the updates of the times. The algorithm is the following:

- Choose one individual at random (let us say k)
- If the state of the individual is 1, which means the individual was infected and removed, and its removal time has been reported the possible move in state is

$$1 \quad \longrightarrow \quad 1$$

We simply update its infection time uniformly in (T_0, r_k), where T_0 is the lower bound for the infection times and r_k is the observed removal time for individual k. The new infection time is accepted with probability

$$A_{1\to 1} = \min\left\{1, \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r})}\right\} \tag{3.19}$$

- If the state of the individual is 0, which means the individual is still susceptible, the possible move in state is

$$0 \quad \longrightarrow \quad 2.$$

We propose to add a pair of infection and removal times. r_k is uniformly proposed in (T_0, T) and s_k is uniformly proposed in (T_0, r_k). In fact, s_k is also uniformly proposed in (T_0, T) conditioned by the fact that $s_k < r_k$. Since for the reverse move we have a probability of 0.5 to propose to delete the pair of removal and infection times added, the acceptance probability is

$$A_{0\to 2} = \min\left\{1, \frac{(T - T_0)(r_k - T_0)}{2} \frac{L(\beta, \gamma, p; \boldsymbol{s}^{(new)}, \boldsymbol{r}^{(new)})}{L(\beta, \gamma, p; \boldsymbol{s}^{(old)}, \boldsymbol{r}^{(old)})}\right\} \tag{3.20}$$

The state of the individual becomes 2 if the infection and removal times have been added.

- If the state of the individual is 2, meaning that a pair of infection and removal times have been added before the current iteration, the possible moves for the states are:

$$\begin{array}{lll} 2 \longrightarrow 0 & & \text{with probability 0.5} \\ \searrow & & \\ \quad\ 2 & & \text{with probability 0.5} \end{array}$$

We propose to delete the pair of infection and removal times with probability 0.5, or update them with the same probability. The acceptance probability for deletion is

$$A_{2\to 0} = \min\left\{1, \frac{2}{(T-T_0)(r_k-T_0)} \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r}^{(new)})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r}^{(old)})}\right\}, \tag{3.21}$$

while the acceptance probability for the update here is

$$A_{2\to 2} = \min\left\{1, \frac{r_k-T_0}{r'_k-T_0} \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r}^{(new)})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r}^{(old)})}\right\} \tag{3.22}$$

where r'_k is the removal time of individual k before the new proposed one r_k.

RJMCMC for the case of incomplete epidemic

This case requires one additional state specifying:

- 3 - Infected not removed

This new state 3 provides more possibilities for moves in the RJMCMC algorithm as in the following:

- Choose an individual at random (let us say k).
- If the state of k is 1, meaning that the individual was infected and its removal time is reported, the state of this individual remains 1 in the algorithm:

$$1 \longrightarrow 1.$$

We update its infection time uniformly in (T_0, r_k). The proposed infection time

is accepted with probability:

$$A_{1\to 1} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}.\right\} \tag{3.23}$$

More efficiently we can make use of the model assumption by proposing the new infection time so that $(r_k - s_k) \sim \text{Exp}(\gamma)$ where s_k is the proposed infection time. In this case the acceptance probability is:

$$A'_{1\to 1} = \min\left\{1, \frac{L(\beta,\gamma;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma;\boldsymbol{s}^{(old)},\boldsymbol{r})} * \frac{\exp\{-\gamma(r_k - s'_k)\}}{\exp\{-\gamma(r_k - s_k)\}}\right\}. \tag{3.24}$$

where s'_k is the current infection time of the individual k.

- If the state of k is 0 (susceptible individual), the possible moves for the states are:

$$\begin{array}{lll} 0 & \longrightarrow & 2 \qquad \text{with probability 0.5} \\ & \searrow & \\ & & 3 \qquad \text{with probability 0.5 .} \end{array}$$

With probability 0.5, propose to add a new infection time or add a pair of infection and removal times.

  - Choose an infection time uniformly in (T_0, T) and add it with probability

$$A_{0\to 3} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{2(T-T_0)}{3}\right\} \tag{3.25}$$

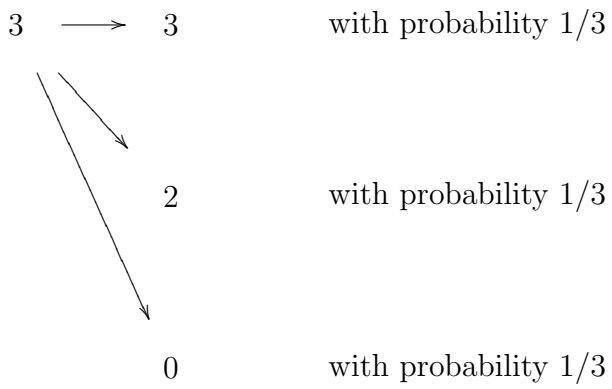
    The 1/3 is the probability of proposing the reverse move, i.e proposing to move from state 3 to state 0 since when in state 3, there are three possible moves and all are made with equal probability. If accepted, the state of the individual becomes 3 which characterises individuals that are infected but not removed before end of the observation period T.

  - Propose a removal time r_k uniformly in (T_0, T) and an infection time s_k in (T_0, r_k) and add the pair with probability

$$A_{0\to 2} = \min\left\{1, \frac{2(T-T_0)(r_k - T_0)}{3}\frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r}^{(new)})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r}^{(old)})}\right\} \tag{3.26}$$

If this move is accepted, the state of the individual k becomes 2 which represents individuals that are infected and removed before T but not reported.

- If state of k is 3, we have the possible moves:



We update the infection time or add a removal time or delete the infection time. With probability 1/3,

  - Update the added infection time by proposing a new infection time uniformly in (T_0, T). The acceptance probability is $A_{3\to3} = A_{1\to1}$.
  - Propose to add a removal time chosen uniformly in (s_k, T) with probability

$$A_{3\to2} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}(T-s_k)\right\} \tag{3.27}$$

    The state of k becomes 2 if the move is accepted.

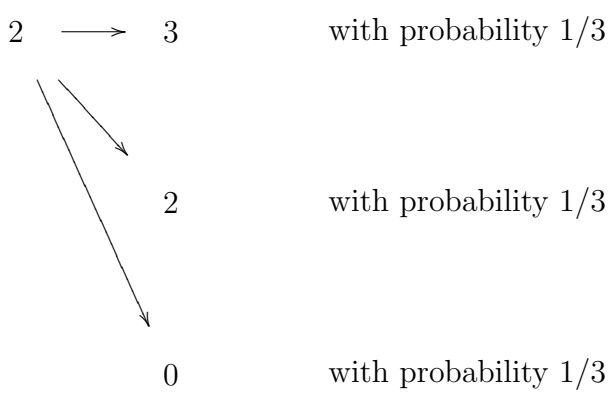
  - Delete the added infection time with probability

$$A_{3\to0} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{3}{2(T-T_0)}\right\} \tag{3.28}$$

    This individual becomes susceptible (state 0) if the move is accepted.

- If state of k is 2 (individual infected and removed before T but not reported),

we can move the state as



With probability 1/3 we either propose to delete the added removal time, or update the couple of infection and removal times, or delete the pair of infection and removal times.

- Delete the removal time added with probability

$$A_{2\to 3} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{1}{T-s_k}\right\} \tag{3.29}$$

  The state becomes 3 when this removal is accepted.

- Update the pair of infection and removal times of k with probability

$$A_{2\to 2} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{r_k - T_0}{r'_k - T_0}\right\} \tag{3.30}$$

  where r'_k is the removal time of individual k before the new proposed one r_k.

- Delete the pair of infection and removal times with probability

$$A_{2\to 0} = \min\left\{1, \frac{L(\beta,\gamma,p;\boldsymbol{s}^{(new)},\boldsymbol{r})}{L(\beta,\gamma,p;\boldsymbol{s}^{(old)},\boldsymbol{r})}\frac{3}{2(T-T_0)(r_k-T_0)}\right\} \tag{3.31}$$

  The state of the individual k becomes 0 if the deletion is accepted.

3.4.3 Update of the reporting probability

The estimation of the probability of reporting requires an update of the unobserved number of removals until time T which is automatically obtained through the event

times updating. In the RJMCMC algorithm in the previous section for the event times updating, the number of removals increases by 1 if a new removal time has been added, decreases by 1 if an added removal time has been removed or remains constant.

We define a beta prior $\mathcal{B}(\alpha_p, \tau_p)$ for p and obtain the beta full conditional posterior distribution

$$p|\boldsymbol{r}, \boldsymbol{s}_{-u}, s_u, \gamma, \beta \sim \mathcal{B}(\alpha_p + n_{rep}, \tau_p + |\mathcal{R}| - n_{rep}). \tag{3.32}$$

This implies that p can be updated through a Gibbs sampling step.

3.5 Application to simulated outbreak data and results

We will first look at a data where we will compare perfect reporting case against under-reporting and apply the RJMCMC algorithm to provide results estimating the under-reporting. We will then make some simulation studies later in Subsection 3.5.5.

3.5.1 Data

We simulate an epidemic based on the Markovian SIR system. The outbreak is taking place in a closed population of $N = 100$ individuals with 99 initially totally susceptible individuals and a single initially infectious case. The parameters for the simulation are $\beta = 0.003$ for the contact rate and $\gamma = 0.1$ for the removal rate. With such parameters, the reproduction number $R_0 = 2.97$ so that an epidemic can happen. We obtain a final size of $n = 93$ individuals ultimately infected after a period of $T = 95$ days. Table 3.1 contains the different specifications for data simulation and the final size of the epidemic for a completed epidemic with perfect reporting.

	N	β	γ	T	n
β	100	0.003	0.1	95	93

Table 3.1: True parameters for data simulation and final size for perfect reporting

Our aim is to study the effect of under-reporting and make inference in the case where we know that under-reporting occurs, at a known rate. We consider a known reporting probability $p \in \{0.4, 0.75, 0.9\}$ and obtain the number of reported cases respectively as $n_{rep} \in \{37, 68, 83\}$. To study the effect of under-reporting, we first make inferences with the data described above assuming perfect reporting. We will then make inferences considering under-reporting.

3.5.2 Comparison between under-reporting and perfect reporting

Non-informative priors for β, γ and p are used at first with parameters $\nu_\beta = \lambda_\beta = \nu_\gamma = \lambda_\gamma = 0.001$ and $\alpha_p = \tau_p = 1$ giving a mean of 1 and variance 1000 for the prior gamma distribution of β and γ; the prior distribution of p is simply $\mathcal{B}(1,1) \equiv \mathcal{U}(0,1)$. Some sensitivity analysis to the prior assumption will be carried later, particularly regarding the sensitivity to the prior of p.

If all the $n = 93$ removal times were observed, MCMC techniques can be applied since we know the dimension of the state space. We can then use a Metropolis-Hastings within Gibbs to estimate the posterior distributions of interest. The parameters β and γ can be updated through Gibbs sampling steps using Equations (3.17) and (3.18). We just need to update the infection times corresponding to all the removal times observed as in Streftaris and Gibson (2004a). The MCMC algorithm leads to the posterior distributions of β and γ summarised in Table 3.2.

	mean	*sd*	2.5%	50%	97.5%
β	0.002842	0.000499	0.002018	0.002788	0.003972
γ	0.0986	0.01945	0.06786	0.0963	0.14333
R_0	2.9011	0.4781	2.0969	2.8601	3.9606

Table 3.2: Posterior estimates of model parameters in the case of complete epidemic with $n = 93$ ultimately infected individuals. All removals are observed and considered in the analysis

We then assume perfect reporting with these number of reported cases of removal times and apply MCMC algorithm to obtain the posterior distributions for β and

γ as summarised in Table 3.3. Clearly, from results in Tables 3.2, 3.3 and Figures

$p = 0.4, \quad n_{rep} = 37$					
	mean	sd	2.5%	50%	97.5%
β	0.001267	0.00035	0.000718	0.00123	0.0021
γ	0.0999	0.0273	0.056	0.097	0.1623
R_0	1.29	0.3134	0.789	1.256	1.998
$p = 0.75, \quad n_{rep} = 68$					
	mean	sd	2.5%	50%	97.5%
β	0.001998	0.00041	0.001328	0.001955	0.0029158
γ	0.1082	0.02274	0.07149	0.1056	0.1602
R_0	1.8584	0.3215	1.3019	1.8325	2.5584
$p = 0.9, \quad n_{rep} = 83$					
	mean	sd	2.5%	50%	97.5%
β	0.00238	0.000461	0.00163	0.00234	0.00342
γ	0.1073	0.0218	0.0719	0.1048	0.1568
R_0	2.2379	0.3689	1.5994	2.2107	3.0518

Table 3.3: Posterior estimates of model parameters in the case of complete epidemic with only reported individuals included in the analysis and assuming perfect reporting ($p = 1$)

3.1, 3.5, 3.9, we can notice that by ignoring the under-reporting in the population, we under-estimate the contact rate β which also results in an underestimation of the reproduction number R_0. Such remark can also be made by looking at the posterior density plots in Figures 3.3, 3.7 and 3.11. An underestimation of R_0 in an epidemic can imply a less or non-efficient measure for eradicating the disease since R_0 is also associated with the proportion of the population that needs to be vaccinated to prevent sustained spread of the epidemic.

It is important to notice that the estimation of β is more accurate and closer to the true parameter value when p increases. This can be seen from the means of the posterior distributions in Table 3.3 where the case $p = 0.4$ has the smallest mean followed by the cases $p = 0.75$ and $p = 0.9$. Therefore, the fewer under-reported cases there exist, the better the estimation of β becomes.

3.5.3 Inference taking into account under-reporting

True value of p : 0.4					
	mean	sd	2.5%	50%	97.5%
β	0.00265	0.00071	0.00134	0.00261	0.00419
γ	0.0991	0.0297	0.055	0.0944	0.169
p	0.439	0.109	0.304	0.417	0.777
n	87.149	13.051	48.000	91.000	100.000
R_0	2.829	1.025	1.288	2.677	5.263
True value of p : 0.75					
	mean	sd	2.5%	50%	97.5%
β	0.00296	0.000624	0.00187	0.00292	0.00433
γ	0.0961	0.0215	0.0623	0.0932	0.1452
p	0.7324	0.0728	0.6106	0.7242	0.9156
n	92.722	6.531	75.000	95.000	100.000
R_0	3.1849	0.9295	1.7672	3.0555	5.3762
True value of p : 0.9					
	mean	sd	2.5%	50%	97.5%
β	0.00283	0.000554	0.00190	0.00278	0.00406
γ	0.0952	0.0210	0.0619	0.0930	0.1441
p	0.885	0.0528	0.783	0.8838	0.9866
n	93.11	4.433	74.000	94.000	100.000
R_0	3.049	0.780	1.8975	3.9218	4.9271

Table 3.4: Posterior estimates in the case of complete epidemic with only reported individuals included in the analysis, and reporting rate taken into account (RJMCMC)

With the different data of $n_{rep} \in \{37, 68, 83\}$ removal times, we now formulate the model including the reporting probability p as in likelihood (3.15). We impute the unobserved event times via the RJMCMC algorithm described in Subsection 3.4.2 to obtain the posterior distributions of the model parameters and the distribution of the unobserved final size as well. The summary statistics are presented in Table 3.4.

As expected, when setting the reporting probability to be constant and equal to 1, the algorithm considers the data as coming from a perfectly reported outbreak without allowing any change of size in the final number of cases. It is merely an MCMC algorithm where for the event times, only the updates of the infection times of the observed removed individuals are made. We then obtain estimates of the posterior distributions given in Table 3.3 with obvious under-estimation of β.

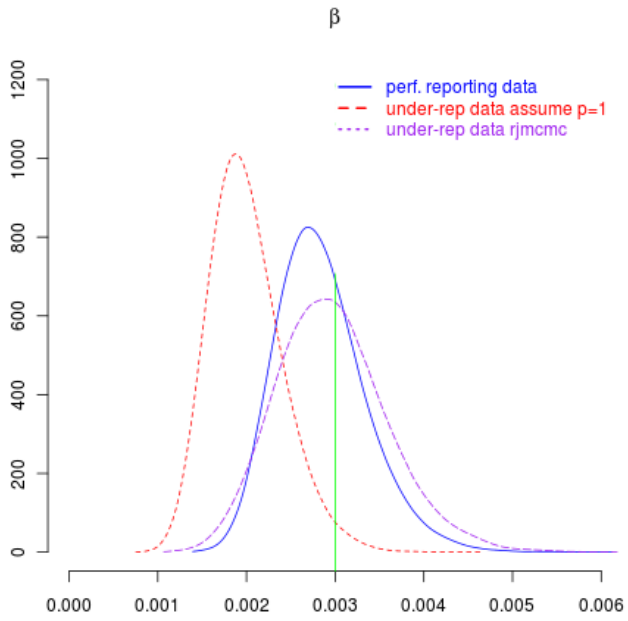


Figure 3.1: Posterior density of β with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.75$

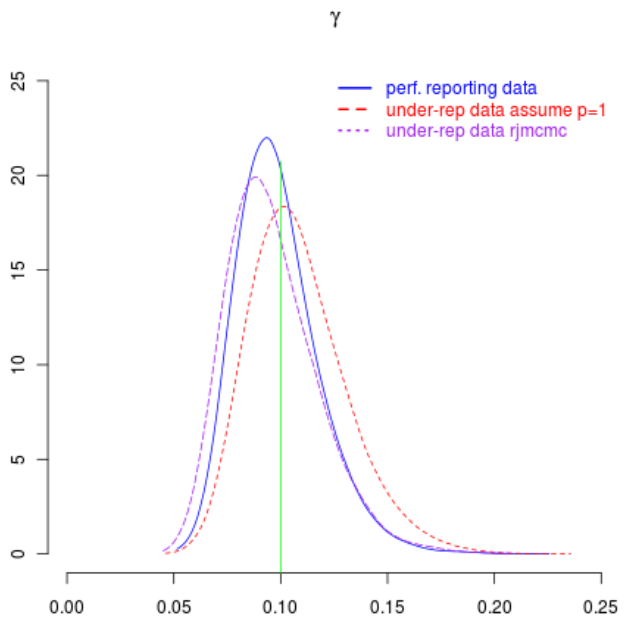


Figure 3.2: Posterior density of γ with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.75$

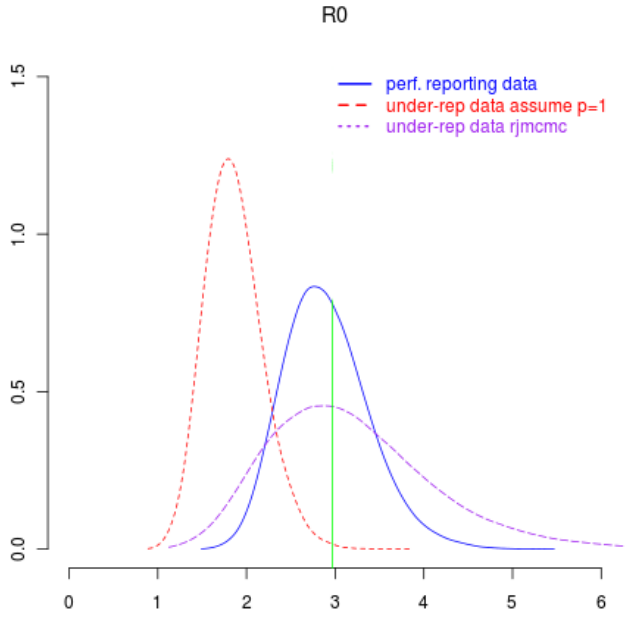


Figure 3.3: Posterior density of R_0 with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.75$

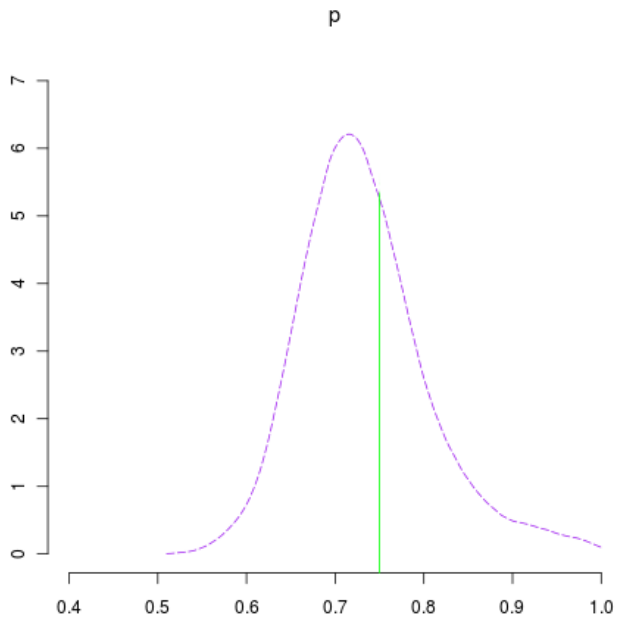


Figure 3.4: Posterior density of p when 68 removal times are reported

Allowing for under-reporting and making full estimation of all the model parameters through the RJMCMC algorithm described in Subsection 3.4.2, we can first notice that our algorithm performs well since our true parameter values are included in their

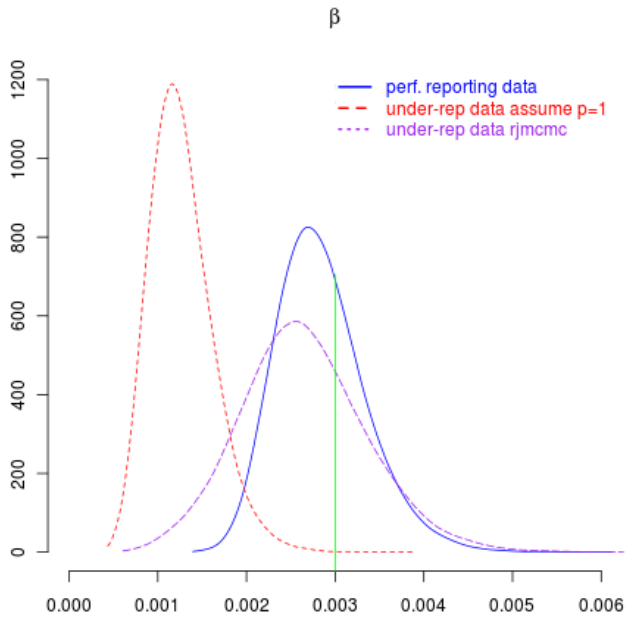


Figure 3.5: Posterior density of β with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.4$

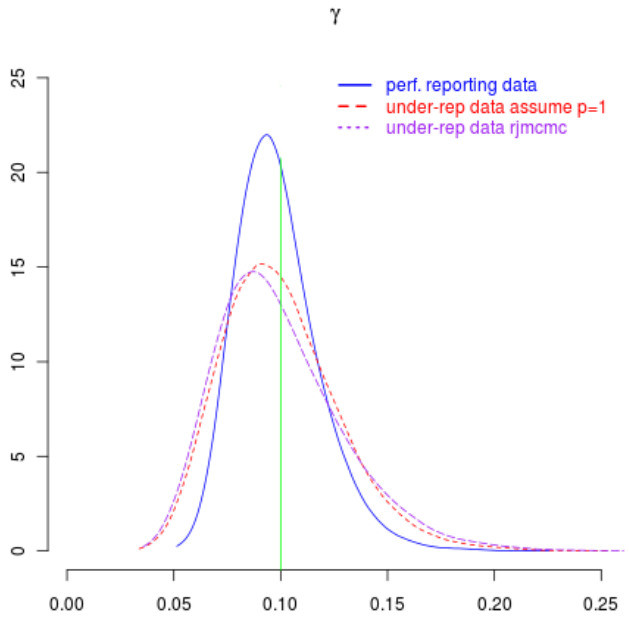


Figure 3.6: Posterior density of γ with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.4$

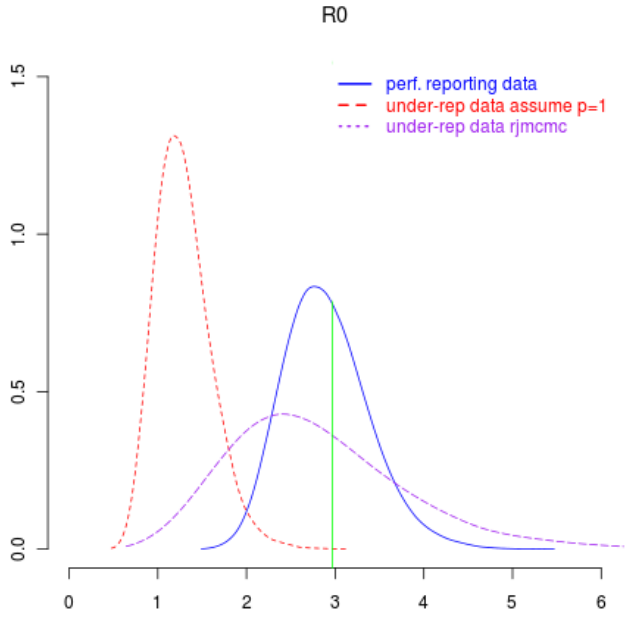


Figure 3.7: Posterior density of R_0 with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.4$

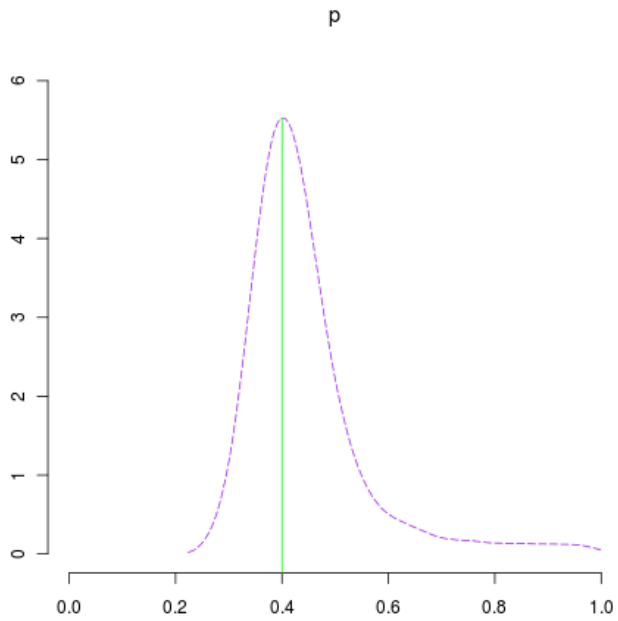


Figure 3.8: Posterior density of p when only 37 removal times are reported

respective credible interval (See Table 3.4 and Figures 3.1- 3.12). In all three cases studied here, we are able to recover the true parameter values used for the simulation of the data.

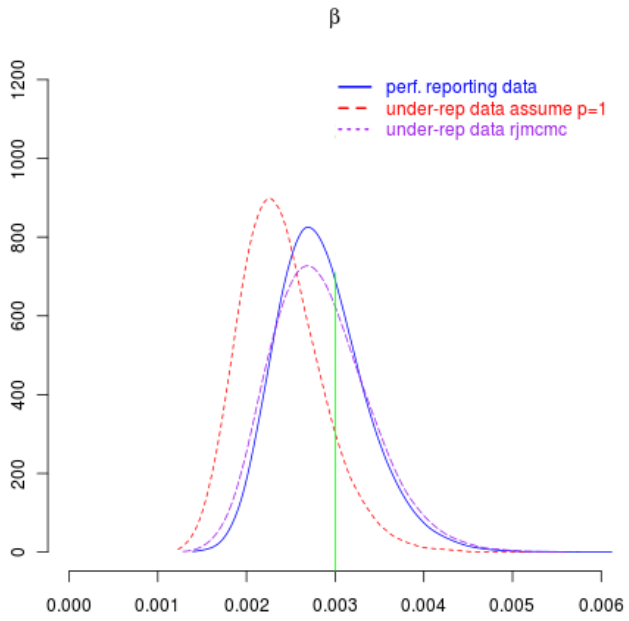


Figure 3.9: Posterior density of β with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.9$

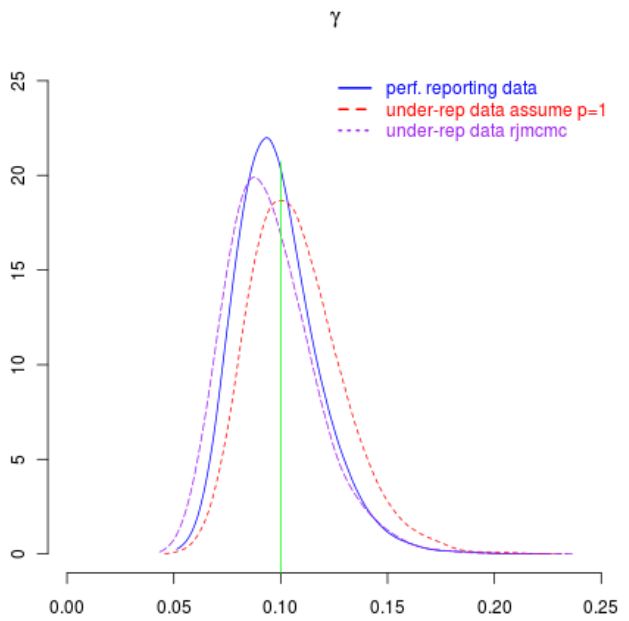


Figure 3.10: Posterior density of γ with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.9$

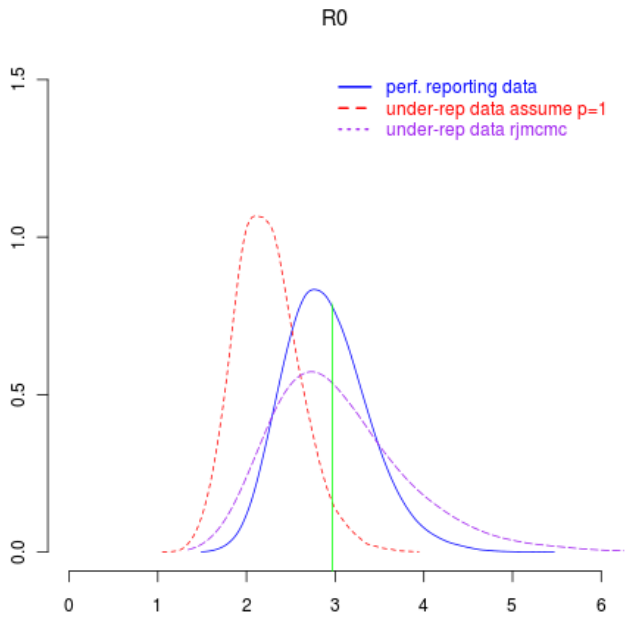


Figure 3.11: Posterior density of R_0 with: full data and perfect reporting (blue solid line); ignored under-reporting (red dashed line); under-reporting taken into account using RJMCMC (purple dotted line) with reported data when $p = 0.9$

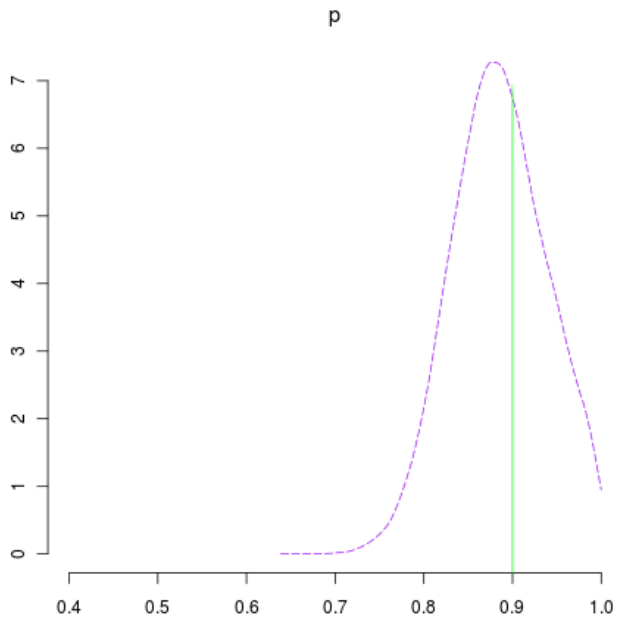
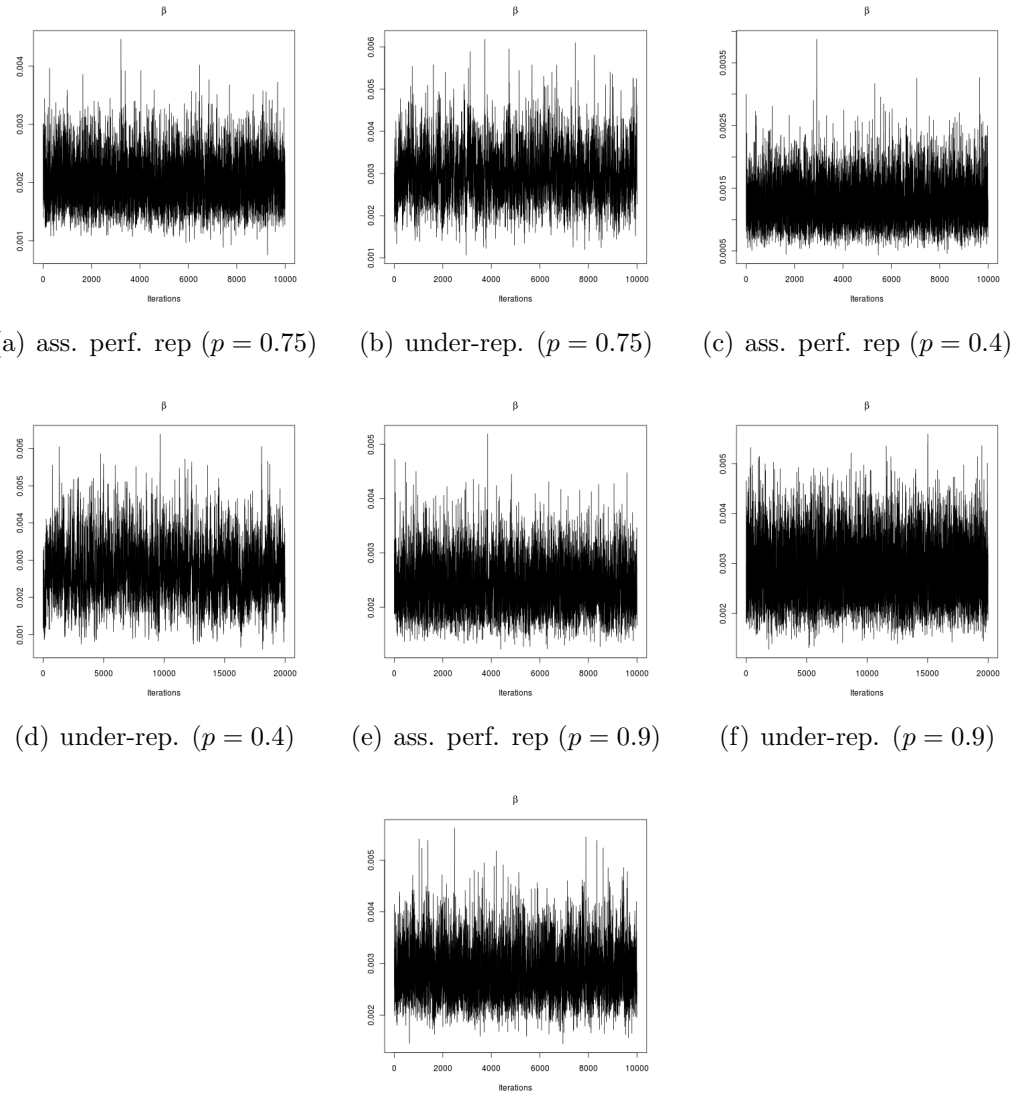


Figure 3.12: Posterior density of p where 83 removal times are reported

In this single simulated data, the estimation of γ is not considerably influenced by the under-reporting, even though the observed final size differs in the different cases considered. The expectation is that under-reporting should be influencing infections



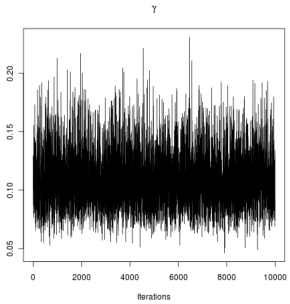
(a) ass. perf. rep ($p = 0.75$) (b) under-rep. ($p = 0.75$) (c) ass. perf. rep ($p = 0.4$)

(d) under-rep. ($p = 0.4$) (e) ass. perf. rep ($p = 0.9$) (f) under-rep. ($p = 0.9$)

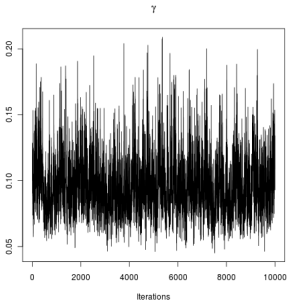
(g) perf. rep.

Figure 3.13: Sample traces for β after a burn-in period of 1000 iterations and a thinning of 10 samples when different reporting probabilities were used to simulate the data

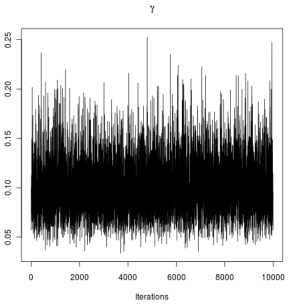
rather than removals. The hidden infections are causing other infections in the population making the true rate of infection been lowered in the estimation in the case of under-reporting. Due to the fact that the estimates of γ are similar in all cases, the posterior densities of β and R_0 appear to have the same shape (see Figures (3.1, 3.3), (3.5, 3.7) and (3.9, 3.11). In general the uncertainty observed in the estimations is related to the amount of information provided in the data. In the case where the number of reported cases is high, the variance in the posterior distributions is smaller compared to the cases of small reported numbers. The case where $p = 0.4$, with only 37 removal times observed, gives the largest variance for the posterior densities fol-



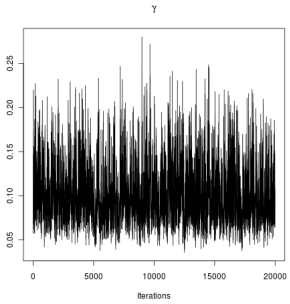
(a) ass. perf. rep ($p = 0.75$)



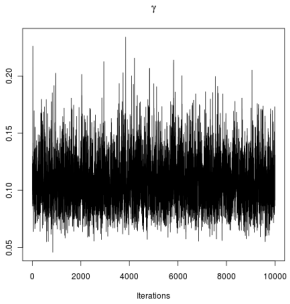
(b) under-rep. ($p = 0.75$)



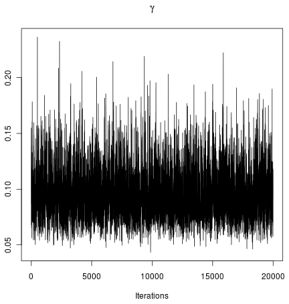
(c) ass. perf. rep ($p = 0.4$)



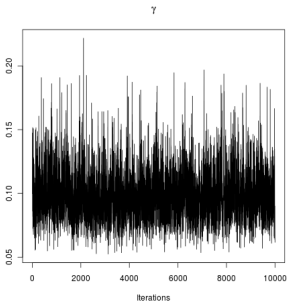
(d) under-rep. ($p = 0.4$)



(e) ass. perf. rep ($p = 0.9$)



(f) under-rep. ($p = 0.9$)

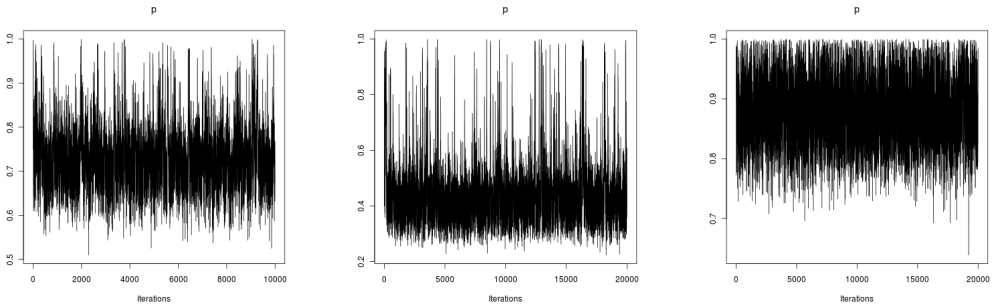


(g) perf. rep.

Figure 3.14: Sample traces for γ after a burn-in period of 1000 iterations and a thinning of 10 samples when different reporting probabilities were used to simulate the data

lowed by the case of 68 removal times ($p = 0.75$) and then the case of 83 removal times ($p = 0.9$). Also the variances in all these cases are higher than the variance in the case of perfect reporting. The distribution of the reporting probability differs in the considered cases. When $p = 0.4$, the estimated posterior density is very long-tailed to the right (Figure 3.8) emphasizing that there is a lot of variability with limited information in the data.

The convergence properties of the Markov chains were explored by looking at the plot of the sample traces of the parameters of interest. Figures 3.13(a)-3.13(g) are the sample traces of β in all the different cases considered for estimation. The chains



(a) Data when $p = 0.75$ (b) Data when $p = 0.4$ (c) Data when $p = 0.9$

Figure 3.15: Sample traces for p after a burn-in period of 1000 iterations and a thinning of 10 samples when different reporting probabilities were used to simulate the data

seems to mix better when the reporting probability is high, meaning that in the case there is a lot of information from the data. When using $p = 0.4$ to simulate the data, the mixing of the chain is slower compared to the case the other cases when p is bigger. This is to be expected as there are a lot of variables treated as auxiliary in the algorithm when p is small. In any case, we do not have evidence of any lack of convergence. The same observations were noticed with the chains for γ in Figures 3.14(a)-3.14(g). The sample traces for p in Figures 3.15(a)-3.15(c) also show us that the mixing of the chains is better when p gets higher. The case $p = 0.4$ displays a long right-hand tail in the distribution as seen in Figure 3.8. In general, there are desirable convergence properties for the algorithms.

3.5.4 Prior sensitivity analysis

Prior sensitivity analysis is conducted on p. We start from assuming less informative prior, moving to very informative prior and also consider a fixed probability of p. We present here the analysis with the data of $n_{rep} = 37$ reported infections ($p = 0.4$). Apart from fixing $p = 0.4$ which corresponds in some sense to a distribution with mean 0.4 and variance 0, we assume successively $\mathcal{U}(0, 1)$, $\mathcal{B}(6, 9)$, $\mathcal{B}(18, 27)$. The corresponding means are respectively $\{0.5, 0.4, 0.4\}$ with respective variances $\left\{\frac{1}{12}, \frac{3}{200}, \frac{3}{575}\right\}$. All the different prior inference cases considered are summarised in Table 3.5.

Very good knowledge about p is equivalent to strong knowledge about the final size of the epidemic n. Therefore, the final size seems to be more dependent on the prior specified. The smaller standard deviation of the prior distribution and of the

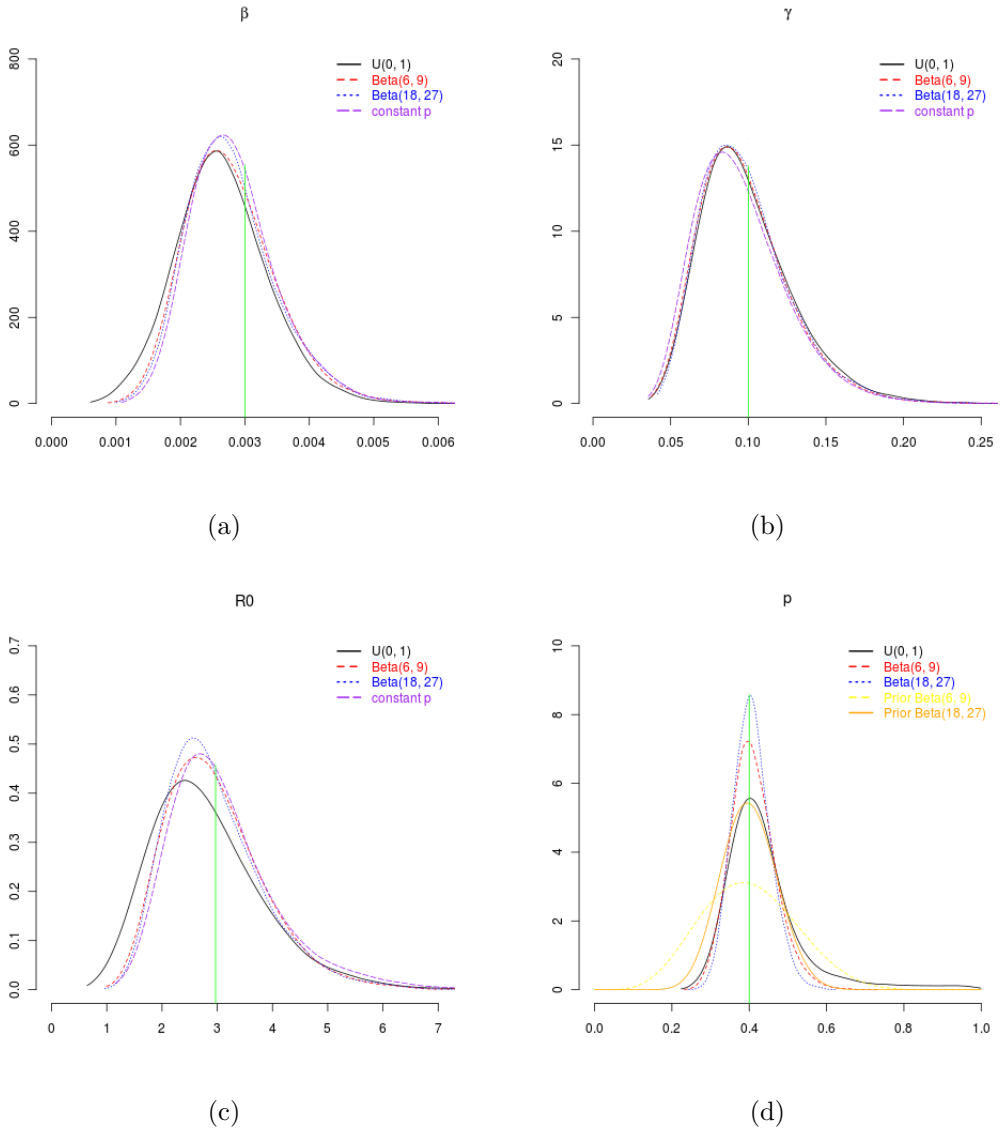


Figure 3.16: Posterior densities of β ((a)), γ ((b)), R_0 ((c)) and p ((d)) assuming different prior distributions for p: $\mathcal{U}(0,1)$ (black solid line); $\mathcal{B}(6,9)$ (red dashed line); $\mathcal{B}(18,27)$ (blue dotted line); and known constant p (purple dashed line)

estimated final size when there is better knowledge of p illustrates this point. It is clear from the graphs in Figure 3.16 that the posterior distributions for β and γ are not very sensitive to the prior on p. The very small trend that we can notice is that the more we know about p, the less variability we have in the estimation of β and therefore R_0. We can conclude that the more informative the prior of p is, the more accurate is the estimation by looking at the mean as point estimate and the standard deviation of the posterior distribution. However, the sensitivity to the prior appears to be quite low.

Fixed $p = 0.4$					
	mean	sd	2.5%	50%	97.5%
β	0.00283	0.00067	0.00175	0.00276	0.00437
γ	0.0954	0.0291	0.0527	0.0908	0.1635
n	92.500	5.902	78.000	94.000	100.000
R_0	3.128	1.024	1.762	2.928	5.732
prior on $p : \mathcal{B}(18, 27)$					
	mean	sd	2.5%	50%	97.5%
β	0.00279	0.000691	0.00166	0.00271	0.00437
γ	0.0983	0.0286	0.055	0.094	0.1674
p	0.406	0.048	0.319	0.404	0.507
n	90.97	7.228	73.000	93.000	100.000
R_0	2.953	0.884	1.649	2.794	5.088
prior on $p : \mathcal{B}(6, 9)$					
	mean	sd	2.5%	50%	97.5%
β	0.00278	0.000694	0.00166	0.00270	0.00434
γ	0.0978	0.0291	0.055	0.0934	0.165
p	0.410	0.0576	0.307	0.405	0.537
n	90.84	7.76	71.000	93.000	100.000
R_0	2.967	0.892	1.637	2.839	5.079

Table 3.5: Summary statistics in the case of complete epidemic with reported individuals using RJMCMC and different prior on p

3.5.5 Simulation study

To be able to capture the variability coming from different datasets, we also carry out a simulation study.

Constant parameters

	mean	sd	2.5%	50%	97.5%
β	0.0034	0.00078	0.0021	0.0033	0.0051
γ	0.100	0.0276	0.0614	0.0957	0.167
p	0.761	0.07068	0.637	0.755	0.911
n	92.35	5.99	78.46	93.82	99.49

Table 3.6: Simulation study in the case of complete epidemic with an average of 70.27 reported individuals and an average of 93.41 ultimately infected; the reporting probability is $p = 0.75$ and non-informative prior are used for p ($Beta(1, 1)$)

For a population of $N = 100$ individuals we set the contact and removal rates to be

respectively $\beta = 0.003$ and $\gamma = 0.1$. The reporting probability is set to be $p = 0.75$ and with non-informative priors for β, γ and p as before, we run a simulation study for 1000 datasets and infer on the parameters, recording the summary statistics of the estimated posterior distributions. In other words, for each simulated data, we run the RJMCMC algorithm for the event times and sample the posterior distributions for β, γ and p using respectively Equations (3.17), (3.18) and (3.32). We consider the summary statistics of the posterior distribution that is then recorded. A second data is simulated and we proceed to inference as before and record again the summary statistics. We repeat the procedure $N_s = 1000$ of times and therefore have 1000 summary statistics coming from different datasets but with the same fixed β, γ and p parameters. An average over the 1000 statistics is computed and put in Table 3.6. The average result, looking at Table 3.6 shows that γ is very well estimated. β is lightly overestimated but fit well in the average credible interval. The average value of p is also very close to the true parameter value and is well within the average credible interval. We are also interested in the coverage property of the credible intervals by estimating the frequency at which the true parameter values fall within their corresponding credible intervals. The contact rate β falls 93.4% of the time in the credible intervals. The reporting probability and the contact rate are included in their credible intervals respectively 93.1% and 96.3% of the time. The estimation overall is very interesting as the true parameter values are recovered.

Same datasets simulation study

To have a better comparison of the different considerations of perfect reporting, under-reporting ignored and under-reporting taken into account, Another simulation study is carried out on the same datasets as we explain in the following. We simulate a reported data for which we know the perfect data if the reporting probability were $p = 1$. Using the perfect data, we estimate the parameters β and γ, and record their summary statistics. We now use the reported data and obtain the posterior distributions of β and γ, treating the reported data as perfect and recording the summary statistics. Further the under-reporting is taken into account on the reported data and RJMCMC is applied for the times with a posterior distribution for β, γ and p obtained. Again the summary statistics of the distribution are kept. We repeat

the scenario a $N_s = 1000$ of times keeping every time the summary statistics in the 3 cases considered. The mean summary statistics are in Tables 3.7 and 3.8 with the different reporting probabilities used.

Perfect reporting					
	mean (MSE)	sd	2.5%	50%	97.5%
β	0.00312 $(3 * 10^{-7})$	0.000534	0.00223	0.00307	0.00432
γ	0.102 $(5 * 10^{-4})$	0.0209	0.069	0.0992	0.151
R_0	3.183 (0.5)	0.566	2.23	3.13	4.44
Under-reporting treated as perfect					
	mean (MSE)	sd	2.5%	50%	97.5%
β	0.00247 $(9 * 10^{-7})$	0.000566	0.00157	0.00240	0.00378
γ	1.38 $(2 * 10^{-3})$	0.0336	0.0851	1.337	2.155
R_0	1.814 (1.35)	3.172	1.272	1.787	2.510
Under-reporting with RJMCMC					
	mean (MSE)	sd	2.5%	50%	97.5%
β	0.00341 $(8.7 * 10^{-7})$	0.000789	0.00212	0.00332	0.00518
γ	0.1003 $(7 * 10^{-4})$	0.0278	0.0608	0.0954	0.167
p	0.759 $(3.6 * 10^{-4})$	0.0710	0.636	0.753	0.912
n	92.459 (27.63954)	6.075	78.127	94.024	99.523
R_0	3.879 (2.34)	1.408	1.890	3.635	7.197

Table 3.7: Simulation-study applied on the same data with $p = 0.75$ where on average $n_{rep} = 70.2$ reported individuals and an average of $n = 93.43$ infected individuals

As expected, the cases of under-reporting considered with the reporting probability $p = 0.75$ are similar from Table 3.6 and the last part of Table 3.7. Also, the two cases of perfect reporting in Tables 3.7 and 3.8 are expected to give similar results since it can simply be viewed as perfect reporting simulation study repeated twice using for each $N_s = 1000$ datasets. The results in both Tables above emphasise the effect of under-reporting pointed out in Subsection 3.5.2. Indeed, when under-reporting exists and it is not taken into account, Tables 3.8 and 3.7 indicate that the estimation of β increases with the reporting probability. In the case where $p = 0.4$, the average credible interval does not contain the true parameter value of β. The estimation of γ seems slightly overestimated when the estimation does not account for under-reporting. By ignoring the under-reporting while it exists, the removal seems to happen a bit faster. But in both cases when $p = 0.4$ and $p = 0.75$, the credible

Perfect reporting					
	mean (MSE)	*sd*	2.5%	50%	97.5%
β	0.00314 $(3 * 10^{-7})$	0.000539	0.00225	0.00309	0.00436
γ	0.102 $(6 * 10^{-4})$	0.0209	0.0693	0.0993	0.151
R_0	3.21 (0.54)	0.578	2.24	3.16	4.49
Under-reporting treated as perfect					
	mean (MSE)	*sd*	2.5%	50%	97.5%
β	0.00168 $(1.9 * 10^{-6})$	0.000465	0.00094	0.00162	0.00274
γ	0.131 $(2 * 10^{-3})$	0.037	0.0729	0.1268	0.2165
R_0	1.298 (2.8)	3.102	0.796	1.263	2.004
Under-reporting with RJMCMC					
	mean (MSE)	*sd*	2.5%	50%	97.5%
β	0.00328 $(8.9 * 10^{-7})$	0.00091	0.00174	0.00319	0.00523
γ	0.111 $(1.5 * 10^{-3})$	0.0359	0.0607	0.105	0.198
p	0.469 $(1.3 * 10^{-2})$	0.111	0.314	0.446	0.739
n	85.108 (202.6744)	12.235	56.638	88.419	98.612
R_0	3.626 (3.07)	1.562	1.453	3.349	7.251

Table 3.8: Simulation-study applied on the same data with $p = 0.4$ where on average $n_{rep} = 37.56$ reported individuals and an average of $n = 93.47$ infected individuals

intervals contain the true parameter value of γ. As a result of the estimation of both β and γ when under-reporting is not accounted for, the basic reproduction is very decreases with p compared to the true value. The variances in the case of RJMCMC are higher.

Sampling of β

Simulation-study with different values for β is carried out as follow. We sample β from a uniform distribution $\beta \sim \mathcal{U}(0.002, 0.0035)$. For each value of β sampled, we simulate an epidemic and estimate the model parameters. We repeat this procedure for $N_s = 1000$ of times and the mean results are computed and summarised in Table 3.9. In fact with such distribution specified for β in the sampling, the mean is 0.00275 with a standard deviation of 0.000433.

Again γ is on average very well estimated from the average mean in Table 3.9. The average mean of the posterior distributions of β is higher than the average sampled value. However, the average credible interval contains the average sampled parameter value. In fact 90.5% of the time, the true value of β is included in the credible interval. The rate of coverage turns out to be respectively 91.2% and 92.7% for γ and p.

	mean	sd	2.5%	50%	97.5%
β	0.00313	0.00069	0.00199	0.00307	0.00466
γ	0.099	0.0267	0.0608	0.0949	0.1633
p	0.748	0.0738	0.617	0.742	0.899
R_0	3.54	1.15	1.86	3.26	6.20
n	90.59	6.27	76.88	91.79	98.82

Table 3.9: Simulation study with β sampled from $\mathcal{U}(0.002, 0.0035)$ with an average of $n_{rep} = 67.6$ reported individuals, $n = 90.00$ infections and an average of $\beta = 0.00275$ sampled giving an average $R_0 = 2.72$

3.6 Conclusions

In this chapter we first presented a general framework for modelling epidemics which allowed us to express explicitly the dynamic probability of the infection and removal processes for threshold models with non-exponential threshold distribution assumption.

We then moved on to consider the Markovian SIR stochastic epidemic model to which we added a reporting process to help us study the effect of case under-reporting. It is clear from our analysis that, in cases where under-reporting in an epidemic is ignored we are led to under-estimation with regard to the extent to which the epidemic could grow. This is obvious from the under-estimation of the reproduction number R_0. To overcome this problem, we presented a RJMCMC algorithm which can help improve estimation as compared to ignoring under-reporting. By applying this algorithm to different cases we noticed that as more removal times are observed through reporting, estimation becomes more accurate. The results were confirmed with simulation studies.

Chapter 4

Estimation of under-reporting using approximations

4.1 Introduction

The general stochastic or Markovian SIR epidemic has been studied in different ways in order to make inference. Recent developments in such studies include the Bayesian approach using MCMC techniques to provide inferences about the model parameters and the basic reproduction number, particularly in the case of partial observations (Gibson and Renshaw, 1998; O'Neill and Becker, 2001; Streftaris and Gibson, 2004a). This inference problem was described in Chapter 2, particularly Subsection 2.8.1, where MCMC turns out to be a powerful tool.

In what follows, we consider the generalised stochastic model as in the previous chapter, and more precisely the general stochastic epidemic in which the infections are reported with probability p independently over individuals. The model can be viewed in two different ways: Either the reporting happens at removal times i.e each of the removal time is independently reported with probability p and we know their corresponding infection times or the infection times are independently reported with probability p and no measure was taken against the reported individuals until the removal happens and the removal time is known. It therefore comes to consider the pair of infection and removal times for each individual are reported with probability p. The problem here is to make inference about the contact and removal rates and the probability of reporting p. The methodology developed throughout this chapter

is well applicable to the generalised stochastic epidemic where different distributions can be used for the infectious period.

The most natural approach would be to derive the likelihood of the model which will require to impute all the unreported infection and removal times, together with all the unobserved infection times for the reported removals in the case of partial observations. Such an approach was explored in the previous chapter and can give interesting results. However, one difficulty associated with this approach is the impracticality of the RJMCMC even for moderately large populations. MCMC methods have been applied to large population sizes (Jewell *et al.*, 2008; Chis-Ster and Ferguson, 2007). Now, due to under-reporting in our model, with all the unknown event times and when there is a large number of individuals that might or might not be infective, RJMCMC becomes impracticable as it will require to impute a very large number of variables and therefore is time consuming.

We are thus led in this chapter to consider approximations which should work well for large population size N. Such approximations will help us make inferences about the parameters of interest using simple MCMC, and thus avoid the computationally intensive change of dimension in RJMCMC. When comparing all the methods later in Subsection 4.5.3, the example with a population size $N = 600$ shows that the RJMCMC algorithm takes more than 36 hours to converge while the approximations require only 4 hours.

The remainder of this chapter is organised as follows. We will first revisit the model and state clearly the approximations. We will then derive an approximated likelihood from which a Bayesian inference method can be implemented to estimate the parameters. The approximations are made using the assumption that exactly a fraction p of the removals (or infections depending on how you look at the model) are reported. In order to allow more uncertainty about p, we will make a suitable correction. Also we will consider an alternative estimation approach in Gibbs-like steps to make inference for the model parameters. We will also develop an alternative and quick approach for point estimation of the parameters by using an appropriate iterative scheme, and we will show some examples with simulated datasets. Another interesting question of interest is to compare the results with full RJMCMC estimation.

4.2 Model and approximations

4.2.1 Description of the model

We consider again the Markovian SIR model or the general stochastic epidemic in which infections and removals (for the same individuals) are reported with probability p, independently over individuals. We recall that the dynamic transition probabilities between states are given by

$$Pr\{S(t+dt) = S(t) - 1\} = \beta S(t) I(t)\, dt + o(\, dt) \tag{4.1}$$

$$Pr\{I(t+dt) = I(t) - 1\} = \gamma I(t)\, dt + o(\, dt). \tag{4.2}$$

For the present Markovian model it clearly does not matter whether we are able to pair the infection and removal times but in general for specific distributions of the infectious lifetime except the exponential, this information would improve inference by making for instance non-centered reparameterisations introduced in Section 2.6.

We assume that each individual, on becoming infected, is reported immediately with probability p independently between infections and its removal is then also observed. Equivalently, the reporting happens at the removal and each removal time is reported with probability p and the corresponding infection time is known for some reason. A Markov specification of the model thus requires 5 categories: $I_o(t)$ and $I_u(t)$ are respectively the total number of reported and unreported infections at time t, $R_o(t)$ and $R_u(t)$ are respectively the total number of reported and unreported removals at time t and $S(t)$ is the total number of susceptibles at time t. Hence we have the total number of infections and removals which satisfy

$$I(t) = I_o(t) + I_u(t), \quad R(t) = R_o(t) + R_u(t) \qquad \text{for all } t \geq 0. \tag{4.3}$$

In terms of sets, the compartments I_o and R_o satisfy $I_o \subseteq I$ and $R_o \subseteq R$. What are therefore actually observed are the processes $(I_o(t), t \geq 0)$ and $(R_o(t), t \geq 0)$ which represent for each compartment a proportion of what would be observed in the case of perfect reporting.

Instead of considering the augmented likelihood in (3.15) with the use of RJMCMC

for inference, we propose here alternative approaches.

4.2.2 Approximations

We henceforth assume a large N and make use of the following approximations. For a given reporting probability p, we assume

$$I_o(t) \approx pI(t) \text{ and } R_o(t) \approx pR(t). \tag{4.4}$$

The approximations in Equations (4.4) correspond to the assumption that exactly a proportion p of the infective cases are reported. Another closely related assumption that we are using is that

$$I_o(t) \sim \text{Bin}(I(t), p). \tag{4.5}$$

For large epidemics, the approximations (4.4) turn out to be accurate. From asymptotic arguments, the more infections there exist, the closer the reported proportion tends to be p times the true number of infections. The question is how valid these approximations when the number of infections are small (beginning of the epidemic) since we make them throughout the evolution of the epidemic. At the early stages of the epidemic, we actually underestimate the variability coming from the reporting process. This will motivate the correction on the reporting probability p as we will see in Section 4.4.

Combining the infection dynamics with the reporting process, we can write

$$\Pr\{ \text{ an infection happens at time } (t+dt) \text{ and it is reported } \} = \beta S(t) p I(t)\, dt + o(\, dt). \tag{4.6}$$

This new process is the same as the one obtained by making use of the random selection property of a Poisson process. Hence we have the approximation

$$\Pr\{ \text{ an infection happens at time } (t+dt) \text{ and it is observed } \} \approx \beta S(t) I_o(t)\, dt + o(\, dt). \tag{4.7}$$

To obtain Equation (4.7), we make use of the approximation $I_o(t) \approx pI(t)$ in Equation (4.6). The approximation in Equation (4.7) simply means that observed infections occur approximately at rate $\beta S(t) I_o(t)$. If $S(t)$, for all $t \geq 0$, was known at each known

infection time, we could make straightforward estimation by deriving an approximated likelihood and proceed to estimation. However, due to the unreported infections, $S(t)$ is unknown and we therefore need to consider further approximations. In the generalised stochastic epidemic model the assumption of closed population implies that the sum of the total number of individuals in each compartment is equal to the population size: $N = S(t) + I(t) + R(t)$. Hence making use of our approximations in (4.4) we approximate $S(t)$ by

$$S(t) \approx N - \frac{I_o(t) + R_o(t)}{p}. \tag{4.8}$$

The two approximations (4.4) and (4.8) put together lead to the probability

$$\begin{aligned} Pr\{ \text{ an infection happens at time } (t+dt) \text{ and it is reported } \} \quad &\approx \\ \beta I_o(t) \left(N - \frac{I_o(t) + R_o(t)}{p} \right) dt + o(\, dt). \end{aligned} \tag{4.9}$$

The new dynamic processes using the approximations above lead us to the approximate likelihood

$$\begin{aligned} L(\beta, \gamma, p; \boldsymbol{s_o}, \boldsymbol{r_o}) \quad \approx \quad & \prod_{i \in \mathcal{I}_{-k}} \beta I_o(s_i^-) \left(N - \frac{I_o(s_i^-) + R_o(s_i^-)}{p} \right) \\ & \exp\left(-\beta \int_{s_k}^{T} I_o(t) \left(N - \frac{I_o(t) + R_o(t)}{p} \right) dt \right) \\ & \prod_{i \in \mathcal{R}} \gamma \exp\left(-\gamma (r_i - s_i) \right) \end{aligned} \tag{4.10}$$

where $\boldsymbol{s_o}$ and $\boldsymbol{r_o}$ are respectively the infection and removal times for reported infected individuals, with k being the first observed infected individual (the times being in increasing order of events).

The likelihood in (4.10) does not involve any augmented variable as compared to the one in (3.15) where augmentation of the data was used. It only involves the available data since it is derived by considering the probability of the occurrence of events that are observed. The estimation of γ is quite straightforward due to the knowledge of the pair of infection and removal times of the reported individuals.

Inference on γ is solely based on the lifetimes of the reported infected individuals. It is expected, using likelihood (4.10), that good knowledge of p will lead to good estimation of β. But in practice, p is also of interest.

The likelihood (4.10) will be used to estimate β and p mainly and inference made using only (4.10) is going to be referred to as method 1 in the remaining of this chapter. But in a sense, the value of p estimated using such approximate likelihood is closer to the proportion that have actually reported, $\hat{p} = R_o(\infty)/R(\infty)$, i.e. the empirical proportion than the true p itself. Therefore, there is further uncertainty associated with the empirical $\hat{p}$ and we will consider allowing for this uncertainty through an appropriate correction later in this chapter. We will also consider a different approach which also uses the likelihood (4.10) but considers a Gibbs sampling steps to provide faster algorithm.

4.3 Inference for β, γ and p using approximate likelihood (Method 1)

4.3.1 Description of Method 1

The likelihood (4.10) is derived using the rate at which observed infections occur and that is a function of β, γ and p. Therefore a direct estimation of all parameters γ, β and p can be made in a Bayesian framework. Here, notice that because the infection times are known, one can easily perform MLE estimation as well. However we adopt the Bayesian framework since it allows flexibility by treating parameters as random variables and incorporation of prior information in the analysis. Other advantages of the Bayesian framework are the fact that we can easily provide information about any quantity of interest that is function of the model parameters (even joint distributional information such as correlations between parameters) and we can assess the accuracy of estimation via credible intervals etc.

The gamma distribution is a conjugate prior for β and γ. If we assume

$$\beta \sim \text{Ga}\left(\alpha_\beta, \nu_\beta\right) \text{ and } \gamma \sim \text{Ga}\left(\alpha_\gamma, \nu_\gamma\right), \tag{4.11}$$

we obtain the conditional posterior distributions

$$\beta|\gamma, p, \boldsymbol{s_o}, \boldsymbol{r_o} \sim \text{Ga}\left(n_{rep} + \alpha_\beta - 1; \int_{s_k}^{T} I_o(t)\left(N - \frac{I_o(t) + R_o(t)}{p}\right) dt + \nu_\beta\right) \quad (4.12)$$

and

$$\gamma|\beta, p, \boldsymbol{s_o}, \boldsymbol{r_o} \sim \text{Ga}\left(n_{rep} + \alpha_\gamma; \sum(r_i - s_i) + \nu_\gamma\right) \quad (4.13)$$

We again assume a beta prior $\mathcal{B}(\alpha_p, \tau_p)$ for p but this time we cannot obtain a beta posterior distribution as in the previous chapter. This is mainly because of the fact that p is involved in the integration part of the likelihood (4.10). However, we can still sample from the posterior distribution of p using Metropolis-Hastings method. As usual it comes naturally to work with the log-likelihood to avoid numerical problems and define the acceptance probability accordingly. The use of random-walk updates for p facilitates the calculation of the acceptance probability

$$Acc = \frac{L^{new} \times (p^{new})^{\alpha_p - 1}(1 - p^{new})^{\tau_p - 1}}{L^{old} \times (p^{old})^{\alpha_p - 1}(1 - p^{old})^{\tau_p - 1}} \quad (4.14)$$

to become

$$\begin{aligned} \log(Acc) \quad &= \log(L^{new}) + (\alpha_p - 1)\log(p^{new}) + (\tau_p - 1)\log(1 - p^{new}) \\ &-(\log(L^{old}) + (\alpha_p - 1)\log(p^{old}) + (\tau_p - 1)\log(1 - p^{old})). \end{aligned} \quad (4.15)$$

The proposed value of p (p^{new}) is accepted if the logarithm of a random variable from a $\mathcal{U}(0, 1)$ is less than $\log(Acc)$ otherwise p is kept to its current value in the chain (p^{old}).

4.3.2 Application using method 1

We apply the methodology above to simulated datasets. Here we recall that the infection and removal times are known for reported infected individuals.

We consider a population of size $N = 600$, and simulate an epidemic based on the two processes described with Equations (4.1) and (4.2). The parameters are chosen to be $\beta = 0.0033$, $\gamma = 1$ (which gives $R_0 \approx 2$) and $p = 0.5$. The choice of the parameters differs from the previous chapter since the population size has increased and we are

only interested in cases where an epidemic can happen. We obtain $n_{rep} = 241$ reported infections out of $K = 482$ ultimately infected as indicated in Table 4.1.

N	β	γ	R_0	p	K
600	0.0033	1.	2	0.5	482

Table 4.1: True parameters for data simulation and final size for perfect reporting

The priors on β and γ are non-informative $\mathrm{Ga}\,(0.001, 0.001)$ distributions. The mean of the prior distribution is 1, the value of γ used for data simulation, but the spread is quite large (the variance is 1000) confirming that γ is non-informative. The prior on p is also non-informative $\mathcal{B}(1,1)$ which is equivalent to a uniform $\mathcal{U}(0,1)$ distribution.

Effect of under-reporting

This was studied in detail in the previous chapter in Subsection 3.5.2. Here we consider a larger population implying a larger dataset with the infection times known for the reported individuals. In the case of perfect reporting, the posterior distributions of β and γ are summarised on Table 4.2. The removal rate γ seems slightly under-estimated. However all the true parameter values are contained in the credible intervals of the posterior estimates.

	mean	sd	2.5%	50%	97.5%
β	0.00323	0.00015	0.00295	0.00322	0.00352
γ	0.9622	0.0439	0.8786	0.9617	1.0500
R_0	2.0124	0.129	1.923	2.007	2.281

Table 4.2: Posterior estimates in the case of complete epidemic with 482 reported individuals (perfect reporting) and infection times known

In the case that we only use the reported times to estimate the parameters by assuming that there is perfect reporting i.e $p = 1$, we have the summary statistics of

posterior estimates shown in Table 4.3.

	mean	sd	2.5%	50%	97.5%
β	0.0022	0.00014	0.00191	0.00217	0.00245
γ	0.9768	0.0627	0.8595	0.9755	1.1036
R_0	1.337	0.121	1.252	1.333	1.589

Table 4.3: Posterior estimates in the case of complete epidemic with 241 reported individuals (assuming perfect reporting $p = 1$) and infection times known

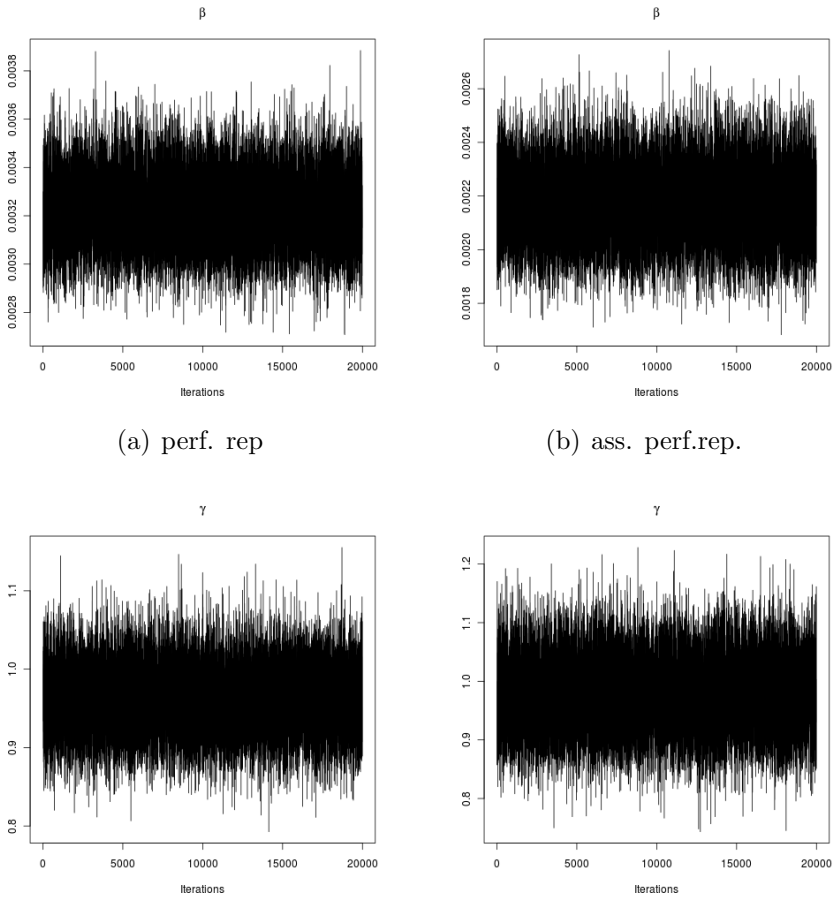
Clearly, β is under-estimated when there is under-reporting but it is not taken into account. The credible interval of the posterior estimate of β does not even contain the true parameter value. This directly implies the under-estimation of the reproduction number R_0 since γ is not influenced by the under-reporting as it is estimated from the $n_{rep} = 241$ reported infectious periods. The densities plotted on Figures 4.2 and 4.4 reflect the comments above. The plots in purple show perfect reporting densities for β and γ while in black are the densities when under-reporting exists but it is not taken into account. We can see how far is the density plot of β when under-reporting exists and it is not accounted for to the density with perfect reporting. The results here are simply based on MCMC algorithm with a Gibbs sampling steps for β and γ since p is assumed to be 1 in each case. Therefore, convergence issues are of no concern as we can see from the sample traces plots in Figure 4.1.

Results from method 1

Using the approximate likelihood (4.10), we obtain the results summarised in Table 4.4.

	mean	sd	2.5%	50%	97.5%
β	0.00323	0.00031	0.00264	0.00322	0.00384
γ	0.9773	0.063	0.8584	0.9762	1.1054
p	0.5148	0.041	0.456	0.5083	0.6117

Table 4.4: Posterior estimates in the case of complete epidemic with 241 reported individuals and infection times known using approximate likelihood



(a) perf. rep (b) ass. perf.rep.

(c) perf. rep (d) ass. perf. rep

Figure 4.1: Sample traces after a burn-in of 1000 iterations and no thinning for β and γ in the case of perfect reporting ((a) and (c)) and when assuming perfect reporting ((b) and (d))

In Figures 4.2 and 4.4, we plot the posterior densities of β and γ respectively using method 1 (in blue) to make inference. For comparison purposes, we also plot the densities in the case p is known (in red) with the case of perfect reporting (in purple) and the case of no under-reporting taken into consideration (assume $p = 1$ in black).

We can see from these plots that the approximations seem to work very well and we are able to recover the true parameter values. Comparing the different densities, in the case where p is known the spread of β is smaller than when p is also estimated. The estimation of γ does not seem to be influenced by the different considerations in Figure 4.4 except in the case of perfect reporting which seems to have smaller variance. This is due to the more data available (482 infectious periods compared to 241 for the other cases). Nevertheless, 241 are adequate in order to have a clear idea about γ, and therefore the loss in the estimation's accuracy is not very significant

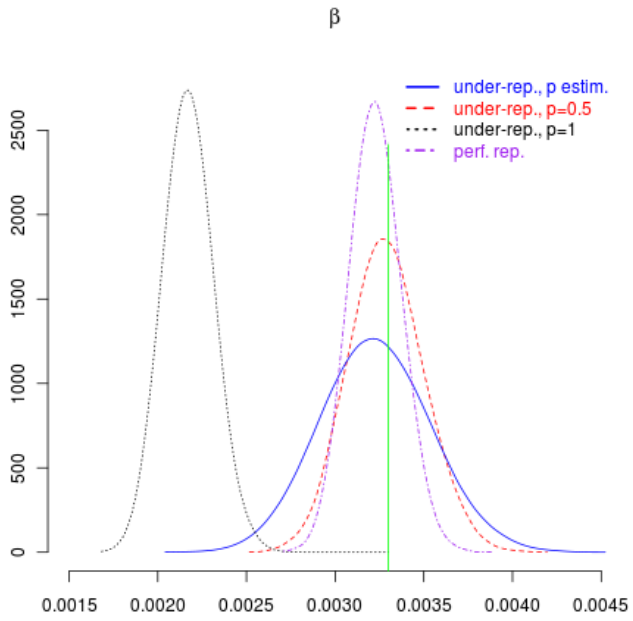


Figure 4.2: Posterior density of β in the case of known infection times: Using approximate likelihood (solid blue line); with perfect reporting (dotted purple line); with imperfect reporting but assumed to be perfect (dotted black line); and with under-reporting probability $p = 0.5$ known (dashed red line)

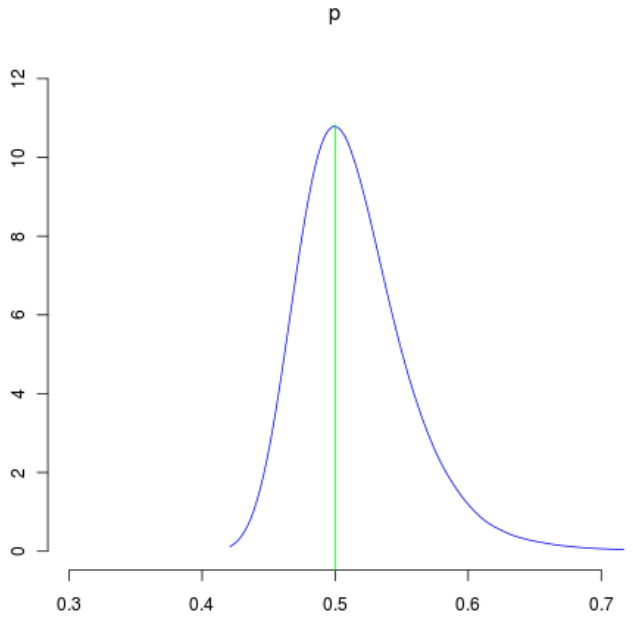


Figure 4.3: Posterior density of p in the case of under-reporting taken into account when using approximate likelihood (Method 1)

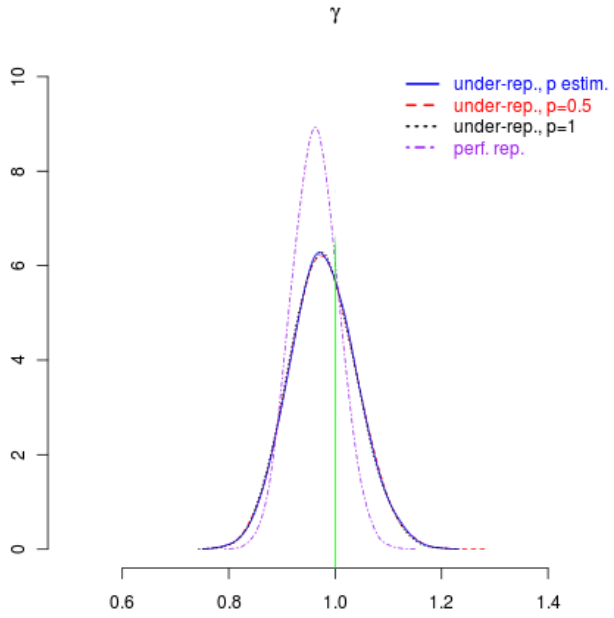


Figure 4.4: Posterior density of γ in the case of known infection times: Using approximate likelihood (solid blue line); with perfect reporting (dotted purple line); with imperfect reporting but assumed to be perfect (dotted black line); and with under-reporting probability $p = 0.5$ known (dashed red line)

compared to the perfect reporting case. We recall that we are working with a large population and $R_0 > 1$. From that perspective, in the remaining of this work we will be interested in β/γ or assume without loss of generality that $\gamma = 1$.

The estimation we have made is based on the approximate likelihood which itself is derived using the approximations (4.4). This approach does not allow the true uncertainty about p but instead provides an estimate of p that may be regarded as the proportion $\hat{p} = R_o(\infty)/R(\infty)$. Hence, because $\hat{p} \simeq p$, there is further uncertainty to associate with p. One possibility would be to add to the posterior variance of $\hat{p}$ a correction factor. But instead of such correction of the variance, we prefer to consider a different method which correctly allows for the uncertainty on p.

4.4 Correction for the estimation of p (Method 2)

4.4.1 Correction with algorithm for inference

Let $K_o = R_o(\infty)$ be the total number of reported infections with the unknown final size of the epidemic being $K = R(\infty)$ (observed plus unobserved) and let $\hat{p} = K_o/K$.

The approximations (4.4) are quite restrictive for the variability of p, as for any time t we have $I_o(t) = pI(t)$ and $R_o(t) = pR(t)$. Thus $\hat{p}$ would have been our estimate of p, had K been observed. Therefore (4.4) can be regarded as an approximation involving $\hat{p}$ (after all this is exactly the case with $t = \infty$) and thus Equation (4.10) is better regarded as a likelihood involving $\hat{p}$ rather than p. To correctly allow for the uncertainty about p, we use our model assumption that the number of reported cases is binomial:

$$K_o \sim \text{Bin}(K, p) \tag{4.16}$$

Interestingly the binomial assumption (4.16) implies that approximately

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{K}\right) \tag{4.17}$$

and because K can be estimated by $K = K_o/p$, we obtain

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p^2(1-p)}{K_o}\right) \tag{4.18}$$

We now have the following for our underlying Bayesian model:

- priors for β, γ (see Distributions (4.11)) and p ($\mathcal{B}(\alpha_p, \tau_p)$)
- $\hat{p} \sim \mathcal{N}\left(p, \frac{p^2(1-p)}{K_o}\right)$,
- an approximate likelihood function $L(\beta, \gamma, \hat{p};\, \boldsymbol{s_o}, \boldsymbol{r_o})$ (i.e. Equation (4.10) with p replaced by $\hat{p}$)

$$\begin{aligned} L(\beta, \gamma, \hat{p}; \boldsymbol{s_o}, \boldsymbol{r_o}) \;\approx\; & \prod_{i \in \mathcal{I}_{-k}} \beta I_o(s_i^-) \left(N - \frac{I_o(s_i^-) + R_o(s_i^-)}{\hat{p}}\right) \\ & \exp\left(-\beta \int_{s_k}^{T} I_o(t) \left(N - \frac{I_o(t) + R_o(t)}{\hat{p}}\right) dt\right) \\ & \prod_{i \in \mathcal{R}} \gamma \exp\left(-\gamma(r_i - s_i)\right) \end{aligned} \tag{4.19}$$

We consider the 3 priors on β, γ and p to be independent as we did in the previous section (4.3) with a gamma prior for β and γ and then a beta prior for p. With $\boldsymbol{s_o}$ and $\boldsymbol{r_o}$ known, the joint posterior density of β, γ, p and $\hat{p}$ is approximately given by

the product

$$\pi_\beta(\beta)\pi_\gamma(\gamma)\pi_p(p)\phi(\hat{p}-p)L(\beta,\gamma,\hat{p};\, \boldsymbol{s_o}, \boldsymbol{r_o}) \tag{4.20}$$

where π_β, π_γ and π_p are the 3 prior densities on β, γ and p respectively, and ϕ is the density of the normal distribution with mean 0 and variance $p^2(1-p)/K_o$.

Thus, treating $\hat{p}$ (essentially K, since K_o is observed) as an unobserved auxiliary variable, we have the basis for MCMC estimation. Let us assume that $\gamma = 1$ for reasons pointed out earlier. In form of pseudo-code, the MCMC is implemented as follows:

Algorithm for correction on p

- Start with initial values for β, $\hat{p}$ and p.
- At each iteration, update our 3 parameters (the auxiliary variable included) each in turn
  - With the current values of p and $\hat{p}$, update β according to its full posterior conditional distribution in (4.12) with p replaced by $\hat{p}$.
  - Given the current values of p and β, update $\hat{p}$ following a random walk scheme with an acceptance probability

$$Acc_{\hat{p}} = \frac{\phi(\hat{p}^{new}-p)L(\beta,\gamma,\hat{p}^{new};\, \boldsymbol{s_o}, \boldsymbol{r_o})}{\phi(\hat{p}^{old}-p)L(\beta,\gamma,\hat{p}^{old};\, \boldsymbol{s_o}, \boldsymbol{r_o})}. \tag{4.21}$$

    Again here a logarithm scale turns out to be a better idea. Therefore our acceptance probability is equivalently

$$\begin{aligned}\log(Acc_{\hat{p}}) &= \log(\phi(\hat{p}^{new}-p)) + \log(L(\beta,\gamma,\hat{p}^{new};\, \boldsymbol{s_o}, \boldsymbol{r_o})) \\ &\quad -(\log(\phi(\hat{p}^{old}-p)) + \log(L(\beta,\gamma,\hat{p}^{old};\, \boldsymbol{s_o}, \boldsymbol{r_o}))\end{aligned} \tag{4.22}$$

  - Given $\hat{p}$ and β (actually this update is independent of β), update p with acceptance probability

$$Acc_p = \frac{\pi_p(p^{new})\phi(\hat{p}-p^{new})}{\pi_p(p^{old})\phi(\hat{p}-p^{old})} \tag{4.23}$$

again with its log-transformation easily obtained as previously.

- Step 2 is repeated until convergence.

The method is then applied to the same data with true parameter values and final size obtained in Table 4.1.

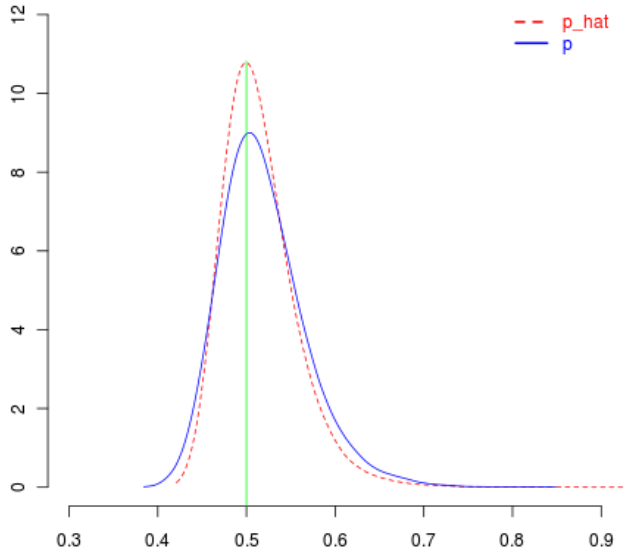
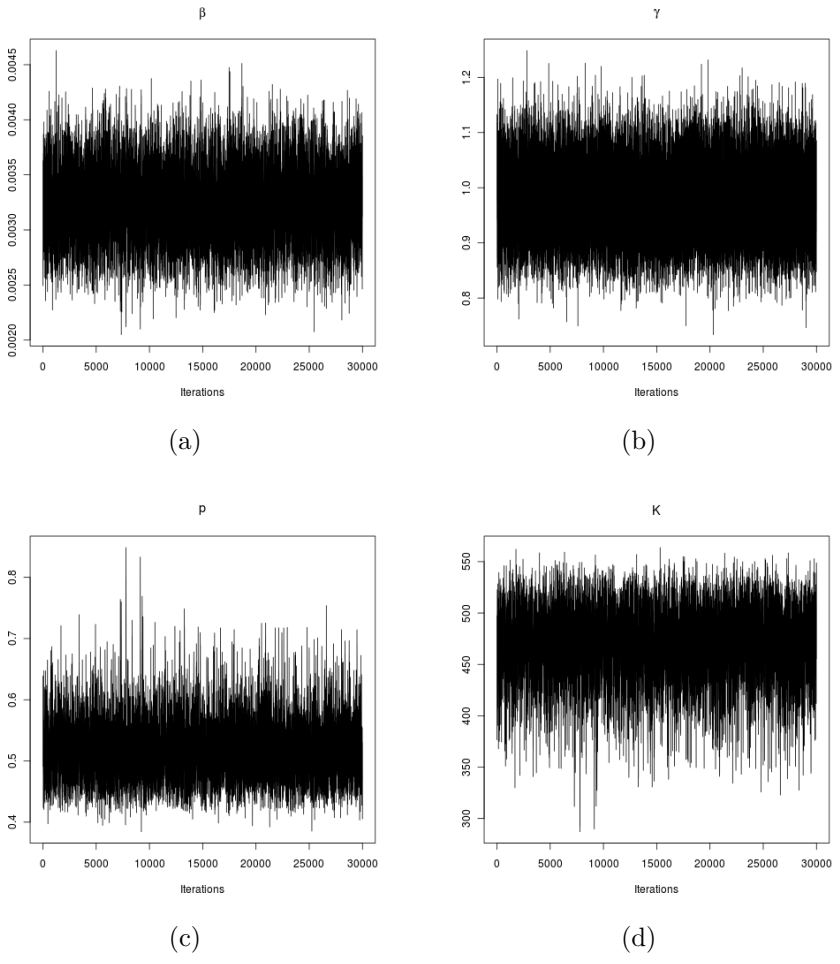


Figure 4.5: Posterior density of $\hat{p}$ (in red) and p (in blue)

	mean	sd	2.5%	50%	97.5%
β	0.00322	0.00031	0.00262	0.00321	0.00384
$\hat{p}$	0.517	0.0424	0.457	0.50983	0.6213
p	0.5187	0.0481	0.4426	0.512	0.631
K	468.93	35.75	387.92	472.75	527.91

Table 4.5: Posterior estimates in the case of complete epidemic with 241 reported individuals and infection times known using approximated likelihood with correction on p

The results for $\hat{p}$ here are non-surprisingly comparable with the results for p in Section 4.3. Such a result in turn makes the estimation of β also closer to the one in Section 4.3. The interest here is to correct the lack of uncertainty on p from the approximations (4.4). It turns out that the means of the posterior distributions of p and $\hat{p}$ are very close but with an increase in the variance for p as we can see in Figure 4.5. A comparison of this result for p with a full RJMCMC update scheme is desirable to assess how well the approximations with the corrections work as we will see in Subsection 4.5.3. We apply different methods to assess the convergence of the Markov chains. One of the methods is to look at the sample traces as plotted in Figure 4.6. The chains mix very well and do not show any evidence of non-convergence. The



(a) (b)

(c) (d)

Figure 4.6: Sample traces after a burn-in period of 1000 iterations and no thinning for the parameters β ((a)), γ ((b)), p ((c)) and K ((d)) when using the appoximate likelihood with correction on p (Method 2)

method using the approximate likelihood (4.10) with correction for p is denoted as method 2 in the remaining of this chapter.

4.4.2 Heuristic justification of the correction on p

Going back to Equation (4.20), we may make the following approximation. For a relatively non-informative prior for p, and so also for $\hat{p}$, the posterior density for $\hat{p}$ is that given by the likelihood function L. Regarding the posterior distributions of $\hat{p}$ and p as being modelled by the random variables $\hat{P}$ and P respectively, we can define the variable X,

$$X = P - \hat{P} \tag{4.24}$$

and let

$$\sigma^2 = p^2(1-p)/K_o. \tag{4.25}$$

Treating σ^2 as a constant, we have

$$X \sim \mathcal{N}(0, \sigma^2) \tag{4.26}$$

Therefore the multiplicative term in (4.20)

$$\phi(\hat{p} - p)L(\beta, \gamma, \hat{p};\, \boldsymbol{s_o}, \boldsymbol{r_o}), \tag{4.27}$$

is the product of the density of X and $\hat{P}$. Hence, integrating over $\hat{p}$ the product gives

$$\int_0^1 \phi(\hat{p} - p)L(\beta, \gamma, \hat{p};\, \boldsymbol{s_o}, \boldsymbol{r_o})\; d\hat{p} \tag{4.28}$$

which, by definition, is the convolution of X and $\hat{P}$ and because X and $\hat{P}$ are independent, the density of $P = \hat{P} + X$. This result suggests that the means of $\hat{P}$ and P are the same and the variance of $\hat{P}$ should be increased by the variance of X to obtain the variance of P. In practice the quantity $\sigma^2 = p^2(1-p)/K_o$ can be estimated by using any reasonable estimate of p.

If we use $\hat{p}$ to estimate σ^2, the uncertainty about p can be estimated by adding to the posterior variance of $\hat{p}$ obtained in Table 4.4 (here p is actually $\hat{p}$), the quantity $\hat{p}^2(1-\hat{p})/K_o$. When using the mean of the posterior density of $\hat{p}$ as an estimate of $\hat{p}$,

we then obtain the variance of P, $Var(P) = 0.041 + \sigma^2 = 0.0022$ which is a standard deviation of $sd(P) = 0.047$. Such value is very close to the one obtained using MCMC algorithm ($sd(P) = 0.048$ as in Table 4.5).

4.5 Inference for β and p using an approximate Gibbs sampling method (Method 3)

We still consider the Markovian SIR epidemic with parameters β and γ and constant probability of reporting p. Parameter γ is assumed known as it can be estimated from the observed infectious lifetimes and here we provide a different method to estimate β and p.

Our alternative approach, which we denote by method 3, is to run a Gibbs sampling approach to estimate both parameters β and p. This requires sampling from the posterior distribution of each parameter conditional on all others. Sampling from the posterior distribution of β given p is straightforward using the approximate density (4.12). To sample from the posterior distribution of p given β, we introduce another auxiliary variable, the final size of the epidemic. In comparison with method 2, method 3 introduces the final size as an auxiliary variable and uses results from the literature to estimate the distribution of the final size. We explore this method in the following subsection.

4.5.1 Estimation of p given β

In what follows, the estimation of p in the case of completed epidemic will require an estimation of the final size $K = R(\infty)$. For an epidemic that is still in progress and observed up to time T, estimation of the total number of infections within the time framework of observation $R(T)$ will be required. With the assumptions (4.4) and (4.8) leading to (4.10), we further assume that all available information about p is contained in the observed final size. Here, knowledge of β is equivalent to knowledge of the reproduction number R_0 (since γ is known). Given the reproduction number, the final size of the epidemic can be estimated as we will see below. Once the final

size is estimated given β, one possibility is to obtain a binomial estimate of p using

$$R_o(\infty) \sim \text{Bin}(K, p) \tag{4.29}$$

where $R_o(\infty) = K_o$ in the case of complete epidemic. If the epidemic is incomplete, we consider the number of reported cases in the time framework of observation $R_o(T)$. As the methodology is briefly described, it remains to be able to estimate the final size of the epidemic K given β.

Final size distribution

Given the contact rate β and the distribution of the infectious period, the distribution of the final size can be obtained solving a system of triangular equations that we need to introduce. Let $\psi(\theta) = \boldsymbol{E}\left[\exp(-\theta\mathbb{I})\right]$ be the moment generating function of the infectious period $\mathbb{I}$, and P_N^k the probability that the final size of the epidemic is equal to k where $0 \leq k \leq n$. Then, P_N^k satisfies the triangular system of equations Andersson and Britton (2000):

$$\sum_{k=0}^{l} \frac{\binom{N-k}{l-k} P_N^k}{\left[\psi\left(\beta(N-l)\right)\right]^{k+a}} = \binom{N}{l}, \quad \text{for } l = 0, \ldots, N. \tag{4.30}$$

In Equation (4.30), N is the initial number of susceptibles in the population and a is the initial number of infectives.

Solutions of (4.30) can lead to negative probability values due to numerical rounding errors. Demiris (2004) discussed that even for moderate population sizes greater than 100, these numerical problems occur with certainty. To avoid numerical problems, the approach proposed in Demiris (2004) is multiple precision arithmetic which is computational costly and time consuming. A quicker way to obtain an estimate of the final size is to use the following Gaussian approximation.

Gaussian approximation

Assume that we have a sequence of Generalised Stochastic Epidemics indexed by the initial susceptible population size N. Let K be the final size of the N^{th} epidemic and τ be the asymptotic proportion of individuals ultimately infected, i.e. $\tau = \lim_{N\to\infty} \frac{K}{N}$. Then τ is almost surely a constant as discussed by Andersson and Britton (2000), and

in the case $R_0 > 1$, τ is the non-trivial solution of the non-linear equation

$$1 - \tau = \exp(-R_0\tau). \tag{4.31}$$

Notice that 0 is always solution of (4.31). A general proof for this result is done by Andersson and Britton (2000). We can interpret this result by looking at the left and right hand sides of the equation separately. The probability of escaping infection, for an individual faced by τ attacks each affecting on average R_0, is equal to 0 occurrence for the Poisson(τR_0) i.e. the right hand of (4.31). On the other hand, the probability of escaping infection is equal to the proportion of initial susceptibles who remain uninfected, i.e the left hand side of (4.31).

If we let $\rho = 1 - \tau$ and $\sigma^2 = var(\mathbb{I})$, then for large N, the distribution of K is approximately Gaussian (Andersson and Britton, 2000):

$$K|R_0 \sim \mathcal{N}\left(\tau N, \frac{N(\rho(1-\rho) + (\beta N)^2\sigma^2\tau\rho^2)}{(1 - \beta N\boldsymbol{E}(\mathbb{I})\rho)^2}\right). \tag{4.32}$$

Demiris (2004) explores this approximation with the multiple precision arithmetic method and validates the approximation for population sizes above 100.

We also consider the Gaussian approximation to the binomial distribution of p in (4.29). We can write that

$$p|K \sim \mathcal{N}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{K}\right) \tag{4.33}$$

where $\hat{p} = K_o/K$.

Given β, we can then sample in turn K from the Gaussian approximation (4.32) and p from the Gaussian approximation (4.33).

Example in the case where β is known

We apply the methodology above to the data described before in Section 4.3 with the true parameter values and final sizes in Table 4.1. We assume β is known and fixed and sample K using (4.32) and then p with (4.33). By repeating this a large number of times, we obtain an approximate distribution for K and p. Table 4.6 contains the summary statistics of the conditional distribution of K and p. Figures 4.7 and 4.8 show the conditional density plots of K and p respectively. The true final size appears

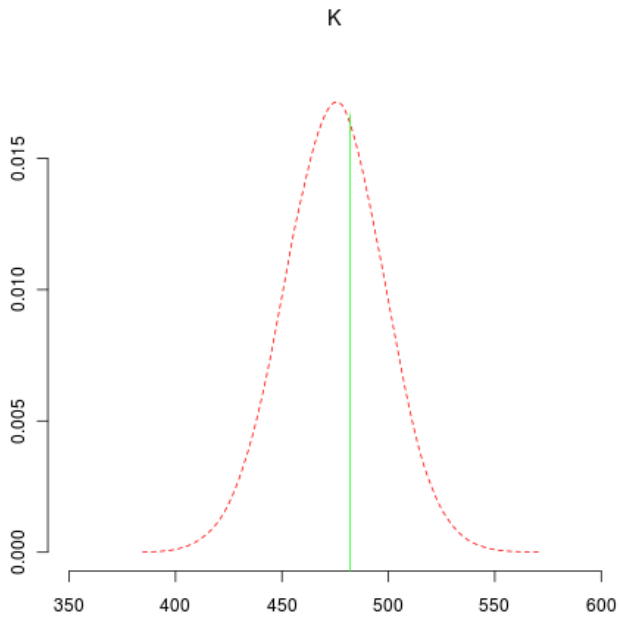


Figure 4.7: Conditional density of K given $\beta = 0.0033$. The green line shows the true final size obtained when simulating the data with perfect reporting assumed

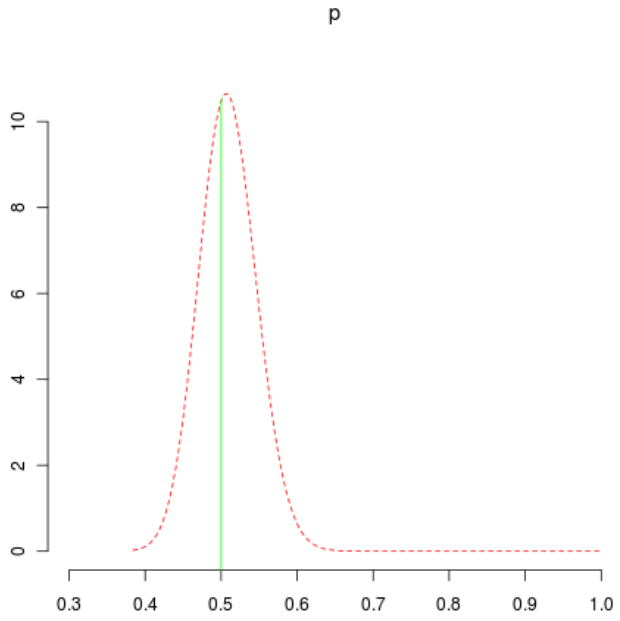


Figure 4.8: Conditional density of p given $\beta = 0.0033$. The green line shows the true reporting probability value $p = 0.5$ when simulating the data

well within the distribution obtained and so does the true value of p. We see from this example that our sampling methods from K and p work very well recovering the true final size and the reporting probability.

	mean	sd	2.5%	50%	97.5%
K	474.61	22.93	429.51	474.73	519.41
p	0.509	0.0338	0.445	0.508	0.5778

Table 4.6: Summary statistics of the predicted final size K and the probability of reporting p in the case of complete epidemic with 241 reported individuals and known value of $\beta = 0.0033$

4.5.2 Approximate Gibbs sampling algorithm

As we have already pointed out, the gamma distribution is a conjugate prior for β using the likelihood in (4.10). With our assumption that all information about p is contained in the final size of the observed epidemic, we introduce the final size $R(\infty) = K$ as an auxiliary variable and we sample from its distribution given β. Then for the obtained sample value of K, we sample from the posterior distribution of p given K and the observed final size K_o. The final size distribution is given by (4.32) conditional on the reproduction number. And finally, given the final size the distribution of p can be simulated using (4.33).

Our aim here is to run a Gibbs sampler for the joint posterior distribution of p and β (or, strictly β/γ if γ is unknown), so that after each half-step this joint distribution remains correct. It is easy to sample from the conditional posterior distribution of β, since assuming a gamma prior we obtain a gamma posterior distribution for β. Thus we condition on β and treat K (the final size) as an auxiliary variable for the present half-step. We sample from the conditional distribution of K, given β, the prior distribution on p and the data. But in fact, for the sampling of K, the prior on p and the data are relatively non-informative since the conditional density of K in (4.32) does not involve p and the reported times. However in the case where nearly every infection is known to have been reported, the sampling of K will be influenced by the prior on p and the data. We retain the non-informative prior for p. Therefore, we can just sample from the conditional distribution of K given β. The only constraint for the sampled value of K is that it must be greater or equal to the reported number of infections ($K \geq K_o$) since the final size of the epidemic cannot be smaller than the reported cases. Now given K (and β which is not directly involved here), we can sample p from its Beta posterior conditional distribution (4.38). The steps of the algorithm are shown below:

Algorithm for approximate Gibbs sampling

1. Given p, sample from the conditional posterior distribution of β

$$\beta|K, p, \boldsymbol{s_o}, \boldsymbol{r_o} \sim \text{Ga}\left(n_{rep} + \alpha_\beta - 1; \int_{s_k}^{T} I_o(t)\left(N - \frac{I_o(t) + R_o(t)}{p}\right) dt + \nu_\beta\right) \tag{4.34}$$

2. Given β, sample the final size K from

$$K|\beta, p, \boldsymbol{s_o}, \boldsymbol{r_o} \sim \mathcal{N}\left(\tau n, \frac{n(\rho(1-\rho) + (\beta n)^2\sigma^2\tau\rho^2)}{(1 - \beta n \boldsymbol{E}(\mathbb{I})\rho)^2}\right) \tag{4.35}$$

   conditional on the fact that $K \geq K_o$.

3. Given K and with $q = K_o/K$, sample the probability of reporting p using

$$p|K, \boldsymbol{s_o}, \boldsymbol{r_o}, \beta \sim \mathcal{N}\left(q, \frac{q(1-q)}{K}\right) \tag{4.36}$$

   The sampling in step (4.36) is a normal approximation to the binomial distribution in (4.29). This can be used in case we define a $\mathcal{U}(0,1)$ prior on p. If we wish to have a more informative prior on p, the term in the binomial approximation (4.29) that contains p is of the form

$$p^{K_o}(1-p)^{K-K_o}. \tag{4.37}$$

   Hence, we can choose a $\mathcal{B}(\alpha_p, \tau_p)$ prior on p and sample p from its conditional posterior distribution

$$p|K, \boldsymbol{s_o}, \boldsymbol{r_o}, \beta \sim \mathcal{B}(K_o + \alpha_p, K - K_o + \tau_p) \tag{4.38}$$

4. With the value of p obtained, go back to step 1.

By repeating the above steps for an adequately large number of times, we can obtain the joint posterior distribution for β and p.

We apply this approximate Gibbs sampling method to the same data described before with the parameters and the final sizes in Table 4.1. The marginal posterior

distributions are summarised in Table 4.7. The density plots of the parameters β,

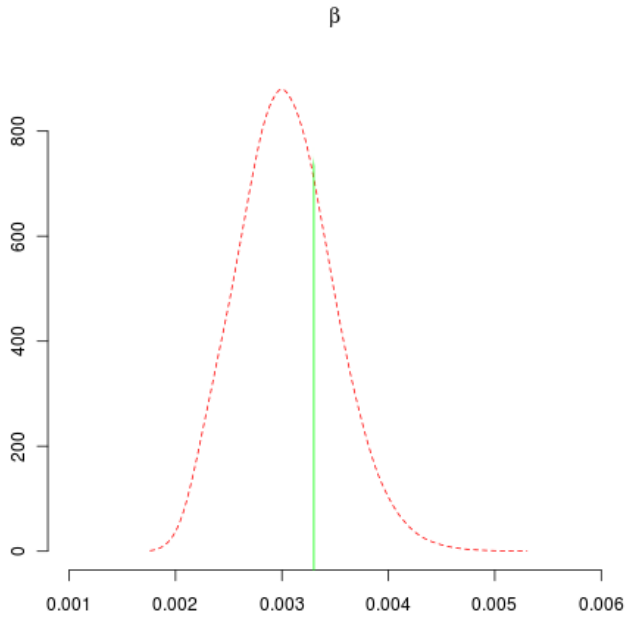


Figure 4.9: Posterior density of β with under-reported data using the approximate Gibbs sampling approach (Method 3)

R_0, K, and p are given in Figures 4.9, 4.10, 4.11 and 4.12 respectively. We can

	mean	*sd*	2.5%	50%	97.5%
β	0.00304	0.0004469	0.00224	0.003024	0.003979
R_0	1.825	0.2681	1.3434	1.8143	2.3874
K	433.809	66.045	278.227	444.6145	534.495
p	0.57	0.1063	0.438	0.5438	0.871

Table 4.7: Posterior estimates of β, the reproduction number R_0, the predicted final size K and the probability of reporting p in the case of complete epidemic with 241 reported individuals using approximate Gibbs sampling approach (Method 3)

see that the true parameter values are well contained within the credible intervals of the obtained posterior densities. We assess the convergence of the Markov chains by applying a number of convergence tests in the literature. One of them consist of

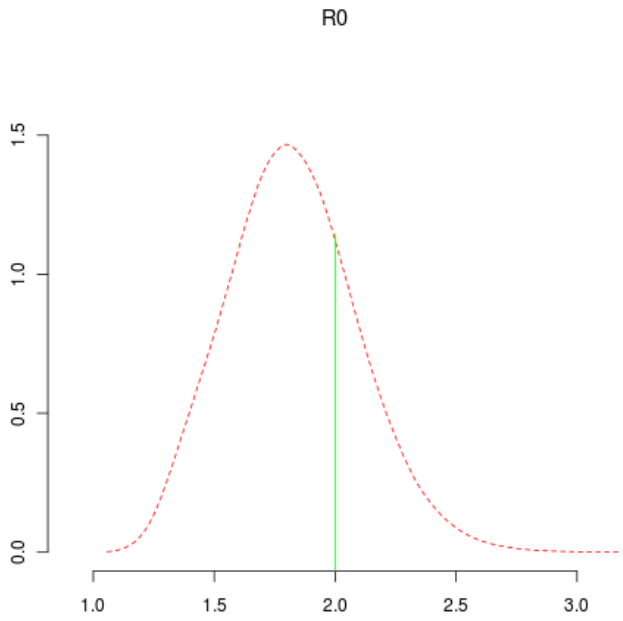


Figure 4.10: Posterior density of R_0 with under-reported data using the approximate Gibbs sampling approach (Method 3)

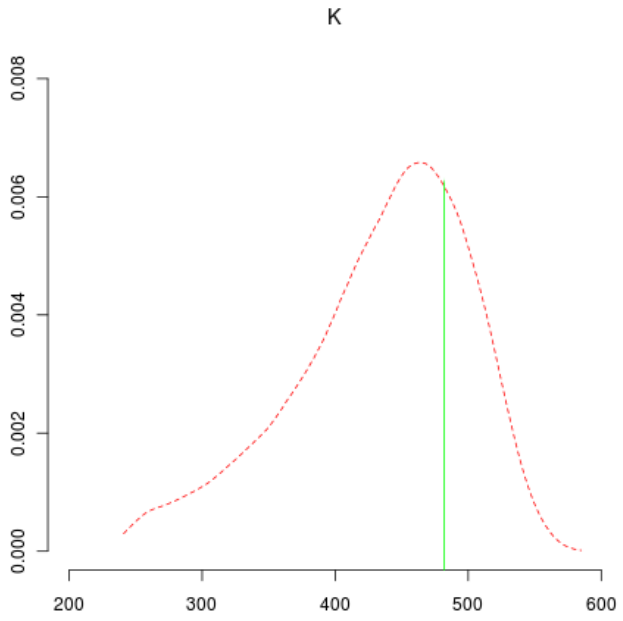


Figure 4.11: Posterior density of K with under-reported data using the approximate Gibbs sampling approach (Method 3)

looking at the sample traces plotted in Figure 4.14. The chains mix well and do not show any particular trend, giving no evidence of non-convergence.

To be able to compare the results for K and p in the cases of known and unknown

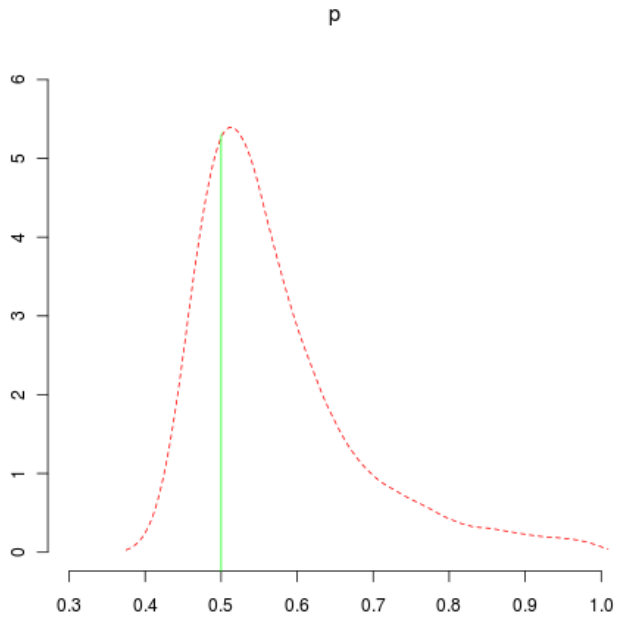
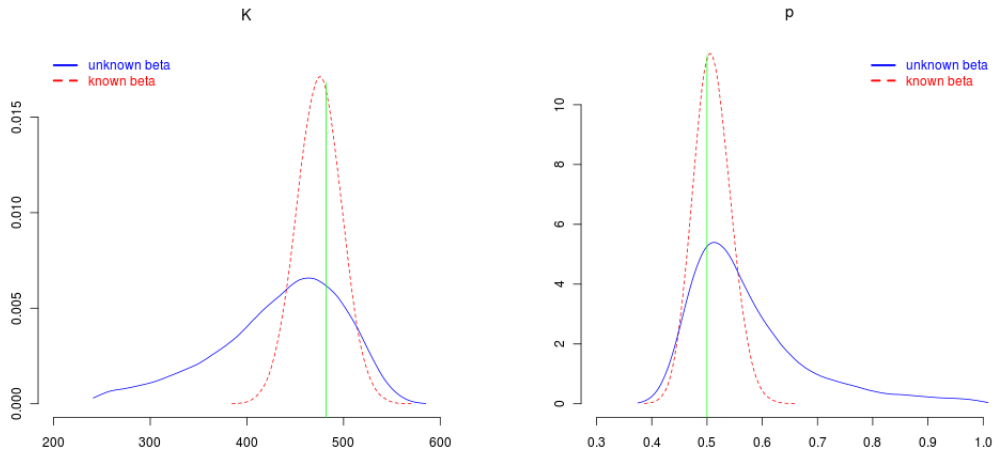


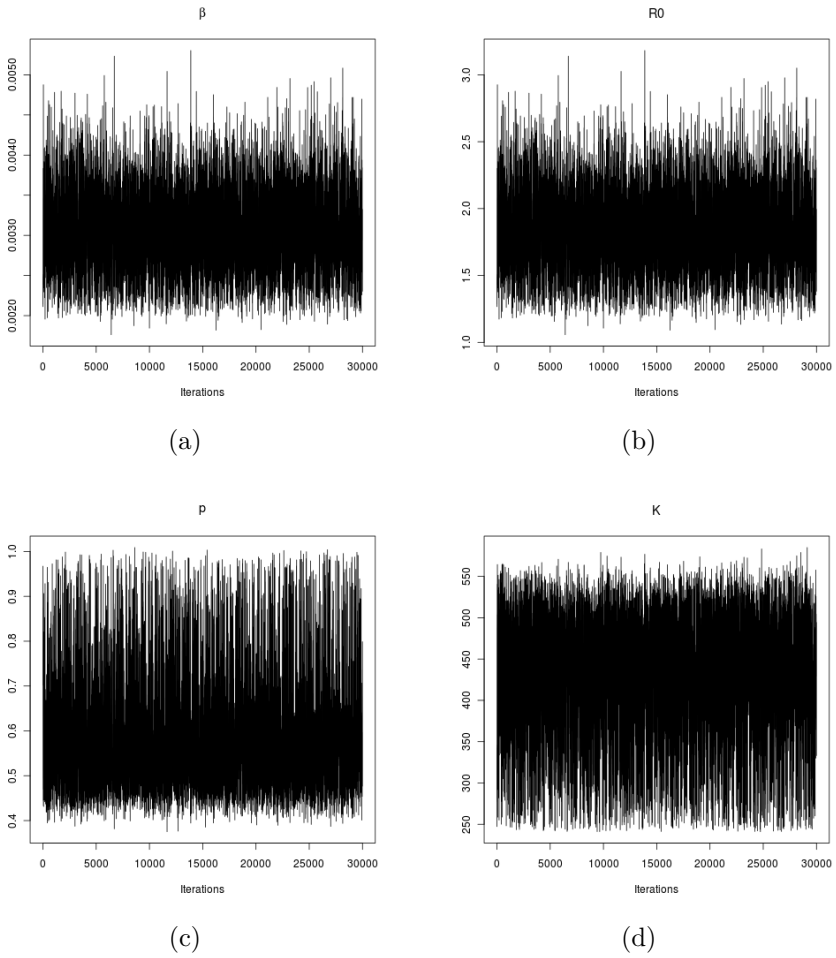
Figure 4.12: Posterior density of p with under-reported data using the approximate Gibbs sampling approach (Method 3)



(a) (b)

Figure 4.13: Posterior density of K ((a)) and p ((b)) in the case where β is known (in red) and when β unknown (in blue)

values of β, we superimpose the plots of the densities in Figures 4.13(a) and 4.13(b) respectively. The distributions of K and p are respectively left-skewed and right-skewed with greater variance when β in not known. This is to be expected since the increase of the uncertainty about β will increase the variance of other parameters. The right-skewness of p is directly correlated with the left-skewness of K since small



(a) (b)

(c) (d)

Figure 4.14: Sample traces after a burn-in period of 1000 iterations and no thinning, for the parameters β ((a)), R_0 ((b)), p ((c)) and K ((d)) when using the appoximate Gibbs sampling approach (Method 3)

values of K lead to very large values of p.

The methodology assumes a large population with an epidemic taking off. A study of how well this methodology works with respect to the size of the population size is also of interest. This is investigated in Subsection 4.6.3.

We have developed inference methodology for epidemics with under-reporting that works well for one simulated data as shown by the results above. All techniques are based on two separate approximations: the first is the approximate likelihood (4.10) while the second is based on the assumption that all information about p in the data comes from the observed final size leading to the binomial distribution in (4.29). The interest now is to compare the results with the first two methods in Sections 4.3 and 4.4 and with a full RJMCMC approach.

4.5.3 Comparison with the full RJMCMC update

The methods in the previous sections involving approximations require a large population size with an epidemic taking off ($R_0 > 1$) to be applicable. At the same time the RJMCMC approach is highly time consuming for large populations. The comparison of the results from the RJMCMC method and the approximate methods would therefore be interesting for a population size that is balanced between a moderately large population size and a relatively small size.

Comparison with a population of size $N = 100$

We therefore apply all the different methods to a population of size $N = 100$, with parameters specified in Table 4.8. The posterior distributions obtained are plotted in

	N	β	γ	R_0	p	K_o	K
β	100	0.003	0.1	3	0.75	64	92

Table 4.8: True parameters for data simulation and final size K for perfect reporting and reported size K_o for a population size $N = 100$

Figures 4.15, 4.16 and 4.17 respectively for β, p and K. Using the different methods, the posterior estimates are summarised in Tables 4.11, 4.9, and 4.10 for methods 2, 3 and RJMCMC respectively.

The densities for β with the RJMCMC and method 1 using directly (4.10) are quite close, while the distribution of β using the Gibbs sampling approach is more spread out. However, the mean of the 3 posterior distributions are very close to each other.

Among the different methods, the Gibbs sampling approach allows more uncertainty about the parameters β and p but give estimates of the posterior mean that are close to the two other methods. Estimation using the likelihood in (4.10), as expected, gives a very narrow posterior distribution for p. As discussed before, it requires a correction on the variance of p to allow properly for the uncertainty about p because of the approximations made to obtain the approximate likelihood. By applying such correction, the posterior distribution of p coincides very well with the exact method

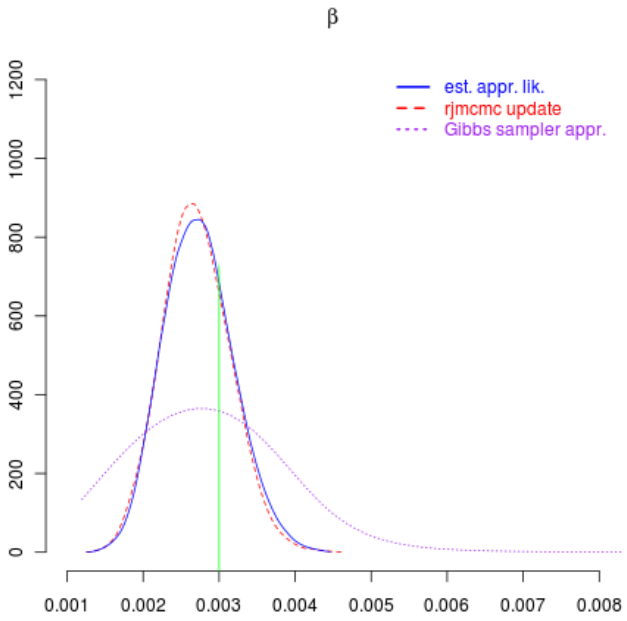


Figure 4.15: Posterior density of β with $N = 100$ data with RJMCMC (red dashed line), direct estimation from (4.10) (blue solid line) and approximate Gibbs sampler approach (purple dotted line)

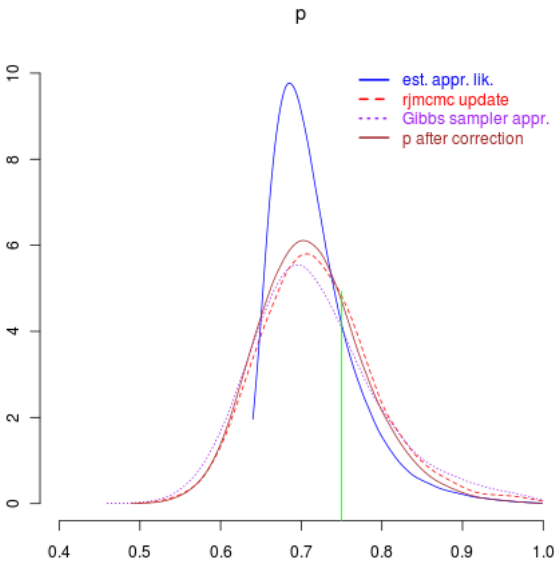


Figure 4.16: Posterior density of p with $N = 100$ data with RJMCMC (red dashed line), meethod 1 (blue solid line), method 2 (brown solid line) and method 3 (purple dotted line)

using RJMCMC as we can see on Figure 4.16. We can see that with a population of size $N = 100$ already, all the methods agree to provide very good estimates of the parameters. It is also interesting to compare the results on a larger epidemic than

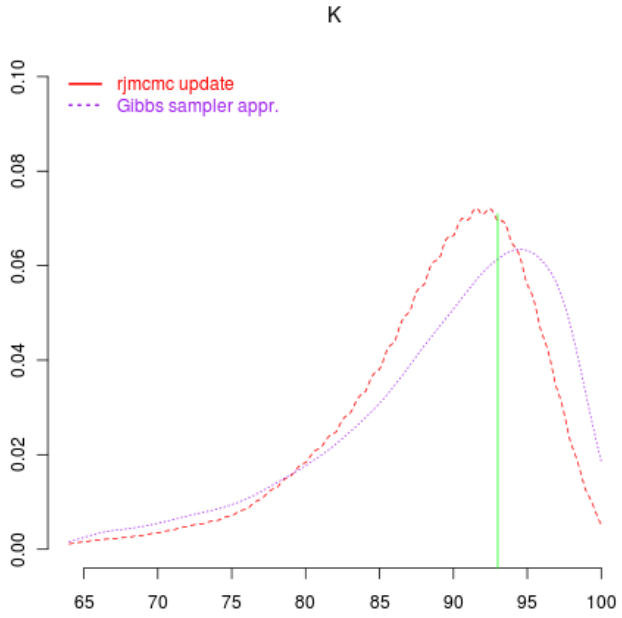


Figure 4.17: Posterior density of K with $N = 100$ data with RJMCMC (red dashed line), and approximate Gibbs sampler approach (purple dotted line)

	mean	sd	2.5%	50%	97.5%
β	0.00283	0.000728	0.00181	0.00272	0.00449
R_0	2.80	0.72	1.79	2.69	4.44
K	89.31	7.52	70.31	91.04	99.05
p	0.72	0.08	0.58	0.71	0.91

Table 4.9: Posterior estimates of β, R_0, K and p in the case of complete epidemic with 64 reported individuals using approximate Gibbs sampling algorithm (Method 3)

	mean	sd	2.5%	50%	97.5%
β	0.0027	0.00044	0.00187	0.00268	0.00361
K	88.74	6.41	72	90	98
p	0.72	0.073	0.59	0.71	0.89

Table 4.10: Posterior estimates of β, K and p in the case of complete epidemic with 64 reported individuals using RJMCMC

	mean	sd	2.5%	50%	97.5%
β	0.00273	0.00046	0.0019	0.00271	0.00367
p	0.72	0.055	0.65	0.704	0.86

Table 4.11: Posterior estimates of β and p in the case of complete epidemic with 64 reported individuals using the approximate likelihood (4.10) (Method 1)

for this epidemic with population size $N = 100$. This is done for a population of size $N = 600$ as follows.

Comparison with a population of size $N = 600$

We move on to apply the 3 different methods we have developed to the data mostly used in this chapter with parameters and sizes in Table 4.1. The posterior distributions obtained under the different methods are summarised in Tables 4.5, 4.7, and 4.12 respectively for the method 2 using the approximate likelihood (4.10) with correction, the Gibbs sampling type approach (method 3) and the RJMCMC. A better demonstration of the results from the 3 methods is shown with the superposition of the posterior densities of the parameters in Figures 4.18, 4.19 and 4.20 for β, γ and K respectively.

	mean	sd	2.5%	50%	97.5%
β	0.003224	0.0003	0.00264	0.003224	0.00382
R_0	1.95	0.205	1.565	1.948	2.369
K	462.63	36.04	381	467	519
p	0.524	0.05	0.447	0.518	0.639

Table 4.12: Posterior estimates of β, K, p and R_0 in the case of complete epidemic with 241 reported individuals using RJMCMC

The posterior estimates of β are quite close in the cases of RJMCMC and the use of likelihood (4.10) as we observed when $N = 100$ datasets. Estimates for the reporting probability p show similar results as for β although the variance is slightly bigger with the RJMCMC method. Again the Gibbs sampler approach provides a wider spread for the posterior distributions of both parameters.

All different methods are able to provide good estimation of the model parameters. Despite the fact that RJMCMC is time consuming it is recommended when the datasets and the population size are not too large. However a quick option while staying within the Bayesian framework would be to use the approximate Gibbs sampler approach or use the approximate likelihood (4.10). Both methods provide acceptable results, with the Gibbs sampler tending to overestimate the variance of the parame-

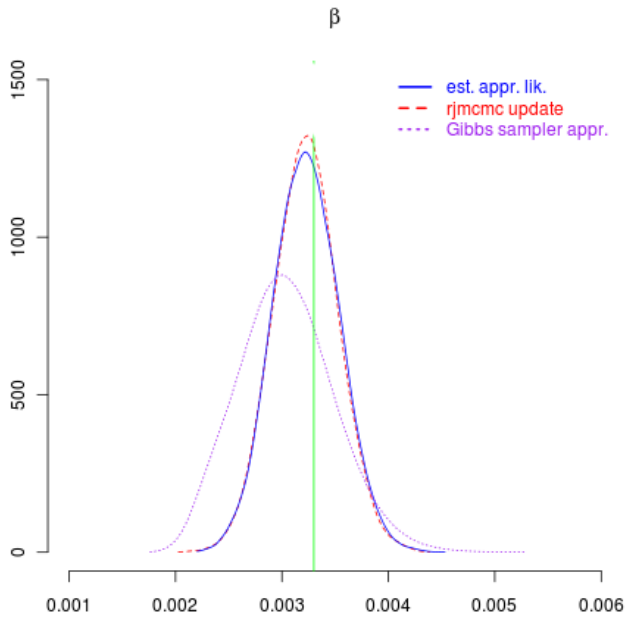


Figure 4.18: Posterior density of β with $N = 600$ data with RJMCMC (red dashed line), direct estimation from (4.10) (blue solid line) and approximate Gibbs sampler approach (purple dotted line)

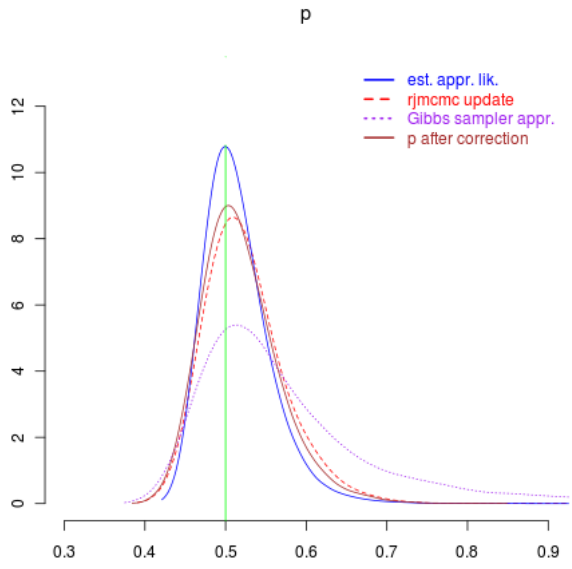


Figure 4.19: Posterior density of p with $N = 600$ data with RJMCMC (red dashed line), direct estimation from (4.10) (blue solid line) with corrected p (brown solid line) and approximate Gibbs sampler approach (purple dotted line)

ters.

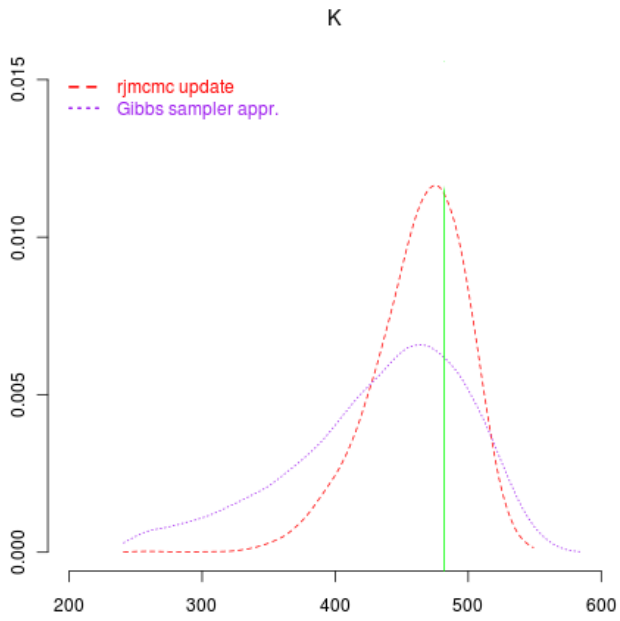


Figure 4.20: Posterior density of K with $N = 600$ data with RJMCMC (red solid line), and approximate Gibbs sampler approach (purple dotted line)

4.6 Simulation Studies

4.6.1 All the model parameters are fixed

We perform a simulation study for the two methods (2 and 3) where the approximate likelihood is used with a correction for the reporting probability (method 2) and Gibbs sampling type approach (method 3). For each of $N_s = 1000$ data sets we make inference for the parameters β, p, R_0 and K using parameter values with population size, true epidemic size and reported size in Table 4.1. The mean, standard deviation, median and credible intervals are monitored for each parameter through the $N_s = 1000$ data sets. A mean is computed for each monitored statistic and the results are shown in Tables 4.13 and 4.14.

We notice that the two methods perform well with a slight advantage for method 2. Indeed, all the mean estimates of the parameters are closer to the true parameter values (Table 4.1) in the case of method 2 than 3 with narrower variances and standard errors. Also, the estimated final size of the epidemic using method 2 is closer to the average simulated true final size than when using method 3. We also look at the rate at which the true parameter values fall within their respective credible intervals. The

	mean (s.e.)	sd	2.5%	50%	97.5%
β	0.003277 (9.7×10^{-6})	0.000317	0.002680	0.0032698	0.003916
K	461.693 (1.38)	36.314	381.538	464.9590	523.173
p	0.51995 (1.5×10^{-3})	0.05193	0.4375	0.5134	0.64132
$\hat{p}$	0.5186 (1.5×10^{-3})	0.04631	0.4517	0.5107	0.632
R_0	1.9759 (5.8×10^{-3})	0.2306	1.555	1.9654	2.4564

Table 4.13: Simulation study result using the approximate likelihood with correction on the reporting probability p (method 2) with an average of $K_o = 236.35$ reported cases. The true average final size simulated is $K = 472.25$

	mean (s.e.)	sd	2.5%	50%	97.5%
β	0.003208 (1.72×10^{-5})	0.000415	0.00247	0.003184	0.004084
K	443.113 (1.92)	54.0912	327.9266	448.2616	532.3333
p	0.56 (3.3×10^{-3})	0.0874	0.4343	0.544	0.767
R_0	1.92167 (0.01)	0.2487	1.479556	1.9071	2.44635

Table 4.14: Simulation study result using the approximate distribution for the final size in the approximate Gibbs sampling approach (method 3) with an average of $K_o = 236.35$ reported cases. The true average final size simulated is $K = 472.25$

coverage rate for β when using method 3 is 83.8%, while it is 94.6% in the case of method 2. For the reporting probability p, the coverage rate turns out to be 85.2% and 96.4% using method 3 and 2 respectively. The two methods rely on large population size and the Gibbs sampling type approach uses an approximate distribution for the final size which also requires a large population size Demiris and O'Neill (2006). This provides one reason why the method with the correction on p performs slightly better than the approximate Gibbs sampling. We note here that the distribution of the final size epidemic is bimodal. The cases where $R_0 <= 1$ i.e epidemics die out quickly are not of interest here. We are only interested in cases where $R_0 > 1$, so that epidemics can occur and the approximate final size distribution is only for the distribution of the high mode of the exact final size distribution.

4.6.2 Different parameter values for β and p

More simulation studies are carried out to evaluate the performance of the methods when the true parameter values vary.

Variation on β only

We first assume that the probability of reporting is fixed, with $p = 0.5$. We then simulate $N_s = 700$ data sets where for each data set the parameter β is sampled from $\beta \sim \mathcal{U}(0.00250, 0.00667)$. This is equivalent to $R_0 \sim \mathcal{U}(1.5, 4)$, since $\gamma = 1$ and is assumed known. R_0 is therefore sampled from a distribution with mean 2.5 and variance 0.52. Again, with non-informative priors on the parameters, summaries of the monitored statistics can be viewed in Table 4.16 for the Gibbs sampling type approach. The results using the method with correction on p are summarised in Table 4.15.

	mean (s.e.)	*sd*	2.5%	50%	97.5%
β	0.00452 (4.5×10^{-5})	0.000365	0.00382	0.00451	0.00525
K	519.89 (2.79)	21.87	472.81	521.61	557.50
p	0.51 (1.8×10^{-3})	0.040	0.442	0.507	0.598
$\hat{p}$	0.509 (1.8×10^{-3})	0.0301	0.462	0.504	0.579
R_0	2.71 (2.6×10^{-2})	0.219	2.29	2.70	3.14

Table 4.15: Simulation study when varying β and keeping p fixed ($p = 0.5$), and using the approximate likelihood with a correction on p method with an average of $K_o = 263.71$ reported cases. The true average final size simulated is $K = 525.91$. The mean of β sampled for the simulation is 0.00453 giving R_0's mean to be 2.72

	mean (s.e.)	*sd*	2.5%	50%	97.5%
β	0.00444 (4.7×10^{-5})	0.000495	0.00353	0.00443	0.00544
K	514.71 (2.82)	31.323	445.80	518.19	564.96
p	0.525 (2.7×10^{-3})	0.051	0.447	0.519	0.641
R_0	2.66 (2.8×10^{-2})	0.296	2.11	2.65	3.25

Table 4.16: Simulation study when varying β, keeping p fixed($p = 0.5$), and using the approximate Gibbs sampling with an average of $K_o = 265.08$ reported cases. The true average final size simulated is $K = 529.02$. The mean of β sampled for the simulation is 0.0046 giving R_0's mean to be 2.759

The results confirm previous finding about method 2 being more accurate despite

	MSE	
	Method 2	Method 3
β	1.3×10^{-7}	2.8×10^{-7}
K	681.428	2768.528
p	0.00241	0.00588
$R0$	0.0462	0.101

Table 4.17: Mean Squared Errors for all the parameters in the case of variation on β using Methods 2 and 3

the fact that both methods perform very well by containing the true parameters in the credible intervals. The coverage rate for β is 93.1% when using method 3 and 95.6% when method 2 is applied. About the reporting probability p, 92% of the time, the true value of p is included in the credible interval when method 3 is applied, and the rate is 94.9% using method 2. In Tables 4.15 and 4.16, we can see that the mean estimates are closer to the parameter values used to estimate the data and that the credible intervals are narrower when using method 2. To compare the two methods more formally, we use the *mean squared error* (MSE)

$$MSE_\theta = \boldsymbol{E}\left(\hat{\theta} - \theta\right)^2 \tag{4.39}$$

as measure; where $\hat{\theta}$ is the mean of the posterior estimate of the parameter θ. For each of the parameters and the final size K, we compute the mean squared error under each method. This can be found in Table 4.17 where the mean squared errors for method 2 are smaller than when method 3 is applied. Actually, the mean squared errors are at least twice in method 3 compared to method 2, suggesting that method 2 is to be preferred. We do the same analysis but this time we vary the reporting probability while keeping the other parameters fixed.

Variation of p only

We carried a simulation study where the value of β was kept fixed ($\beta = 0.0033$). At each data simulation, the value of the reporting probability was sampled from a uniform distribution, namely $p \sim \mathcal{U}(p_{min}, p_{max})$. By setting $p_{min} = 0.3$ and $p_{max} = 0.7$, we run $N_s = 1000$ simulations for data and carry out inference for each of them. An average summary of the results are in Tables 4.19 for method 2 and 4.18 for method 3. The mean value of p to be sampled for simulating the data is $\bar{p} = 0.5$.

From this perspective, the results of Tables 4.19 and 4.18 are comparable with the results respectively in Tables 4.14 and 4.13 where it turns out that the differences are negligible.

	mean (s.e.)	sd	2.5%	50%	97.5%
β	0.00329 (1.2×10^{-5})	0.000327	0.00267	0.00328	0.00395
K	462.23 (1.74)	37.29	379.70	465.77	524.97
p	0.51 (4.4×10^{-3})	0.051	0.43	0.504	0.63
$\hat{p}$	0.51 (4.4×10^{-3})	0.046	0.44	0.50	0.62
R_0	1.980 (8.1×10^{-3})	0.24	1.55	1.97	2.48

Table 4.18: Simulation study when varying p and keeping β fixed ($\beta = 0.0033$), and using the approximate likelihood with a correction on p method with an average of $K_o = 233.00$ reported individuals. The true average final size is $K = 473.33$. The mean of p sampled for data simulation is 0.493

	mean (s.e.)	sd	2.5%	50%	97.5%
β	0.003205 (2.1×10^{-5})	0.0004172	0.00247	0.003178	0.004090
K	442.29 (2.42)	54.625	327.33	447.14	533.16
p	0.56 (5.4×10^{-3})	0.0871	0.435	0.544	0.767
R_0	1.92 (1.3×10^{-2})	0.250	1.480	1.904	2.450

Table 4.19: Simulation study when varying p and keeping β fixed ($\beta = 0.0033$), and using the approximate likelihood with a correction on p method with an average of $K_o = 238.32$ reported individuals. The true average final size is $K = 474.54$. The mean of p sampled for data simulation is 0.501

	MSE	
	Method2	Method3
β	1.01×10^{-7}	3.18×10^{-7}
K	1507.627	6364.994
p	1.4×10^{-2}	2.8×10^{-3}
$R0$	0.046	0.12

Table 4.20: Mean Squared Errors for all the parameters in the case of variation on p using Methods 2 and 3

Again here, by looking at Tables 4.14 and 4.13, we can conclude that even with varying values of p, we can very well estimate the model parameters and therefore recover the loss due to under-reporting. Also the true value of β is included in the credible interval 84.2% of the time, when method 3 is used, and 95.6% for method 2. The true reporting probability is contained in the credible interval 86.6% of the time, when we apply method 3 and 95.4% using method 2. In terms of comparisons, method 2 seems to perform better than method 3 as the tables indicate more closer mean estimates to the true parameter values. To compare formally the performance of the two methods, we compute the mean squared error of all the parameters in Table 4.20. The mean squared errors in the case of method 2 are smaller than the case of method 3. This confirms the fact that method 2 is giving more accurate results than method 3 even though both methods are very interesting and give very useful results.

4.6.3 Simulation studies with varying population size

Method 3 performs less accurately than method 2 with all the different datasets considered when the comparisons are made. We then carry out a simulation study with $N_s = 1000$ datasets on a larger population of size $N = 10000$ to compare how well method 3 performs as the size of the population increases. With parameters $\beta = 0.0002$, $\gamma = 1$, and $p = 0.5$, the results for the simulation are summarised in Table 4.21 for method 3. The reporting probability p is kept fixed with the case where the

	mean (s.e.)	sd	2.5%	50%	97.5%
β	0.000199 (2.87×10^{-7})	0.00000628	0.000187	0.000199	0.000211
K	7913.538 (8.52)	201.37	7493.034	7922.714	8281.307
p	0.504 (6.7×10^{-4})	0.0144	0.479	0.504	0.535
R_0	1.99 (2.12)	0.063	1.869	1.99	2.11

Table 4.21: Simulation study result using the approximate Gibbs sampling method with fixed parameters $\beta = 0.0002$, $p = 0.5$ and an average of $K_o = 3984.359$ reported cases. The true average final size is $K = 7965.40$

population size was chosen to be $N = 600$. It is clear with this new simulation study that the method performs very well with larger population size assuming the epidemic

does not die out very quickly at early stages. Indeed, estimate of p is closer to the true parameter value when the population size is large since with $N = 10000$, the mean of p is 0.504 with a standard deviation equal to 0.0144, while with $N = 600$, the mean of p is 0.56 with a standard deviation of 0.0874. It is confirmed by the standard errors that are much smaller with larger population size.

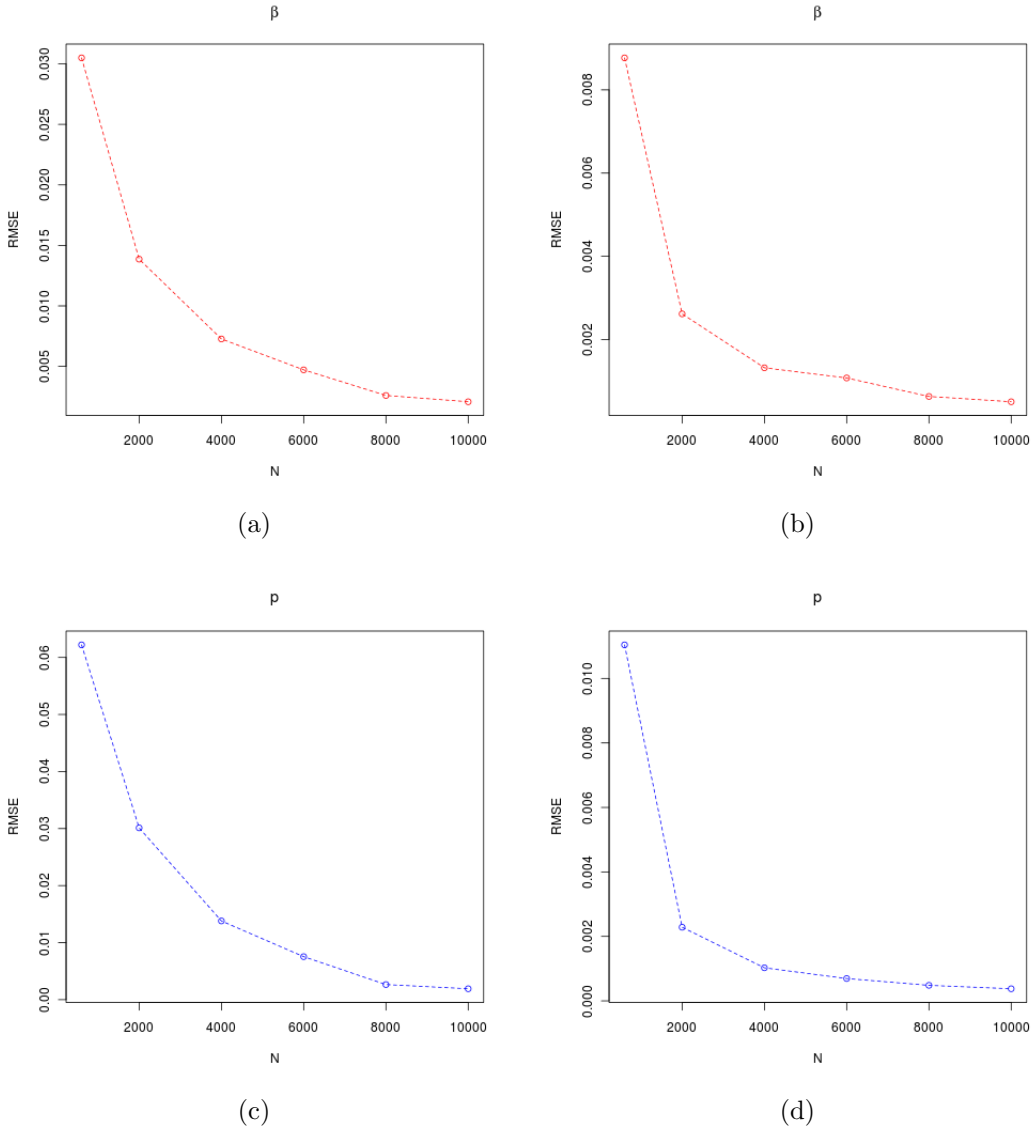


Figure 4.21: Plots of the relative mean squared error of β and p as a function of the population size. (a) and (c) are plotted after using the Gibbs sampling approach while (b) and (d) are plotted after the approximation with correction on p method. Note different scale on the vertical axis

We now explore how methods 2 and 3 perform with different population sizes. We

run simulations with different population sizes by choosing the contact rate β such that the reproduction number R_0 is unchanged ($R_0 \approx 2$). With the different population sizes $N \in \{600, 2000, 4000, 6000, 8000, 10000\}$, the contact rate β was chosen to be respectively to be $\beta \in \{0.0033, 0.001, 0.0005, 0.00033, 0.00025\}$. For each of the simulation studies, we consider the mean of the posterior distributions as a point estimate and compute the relative mean squared error (RMSE) of the parameters β and p. When applying method 3, we plot the RMSE as a function of the population size N in Figures 4.21(a) and 4.21(c) respectively for β and p. The plot of the RMSE when using method 2 for β and p are in Figures 4.21(b) and 4.21(d) respectively. There is a decrease in the relative mean squared error as the population size increases. The relative mean squared error is therefore inversely proportional to the population size. The asymptotic assumption holds for the two methods: the bigger the population size is, the more accurate the estimations are. However, when looking at the 4 figures 4.21(a), 4.21(c), 4.21(b) and 4.21(d), we can again notice that method 2, where the approximate likelihood (4.10) is used with a correction on the reporting probability p, perform better than the Gibbs sampling method (method 3). The relative mean squared error is smaller when using method 2 than in the case of method 3. This was concluded looking at Tables 4.13 and 4.14, and is now confirmed by the relative mean squared errors.

All the methods performed are based on Bayesian methodology. However, the approximate likelihood (4.10) can be used to obtain MLE estimates and we use it below to show how we can quickly obtain point estimates for the model parameters.

4.7 Iterative scheme for point estimate

Maximum Likelihood Estimation (MLE) can be applied here to obtain point estimates for parameters β and p.

A straightforward way to estimate the final size is to use the result in Equation (4.31) that we recall

$$1 - \tau = \exp(-R_0\tau), \tag{4.40}$$

where τ is the proportion of the population that becomes ultimately infected. From τ a point estimate of the final size is $K = \tau \times N$. We repeat the steps below until

convergence; by convergence we mean the relative difference between two successive estimates of β and p is negligible. The steps of the iterative scheme are as follows:

Iterative scheme algorithm

1. Start with $p = 1$ and find the MLE of β using the approximate likelihood (4.10)

2. With the MLE of β:

   - compute and estimate the reproduction number $R_0 = \beta \times (N - 1)$
   - Find the proportion of infected solving $1 - \tau = \exp(-R_0\tau)$
   - The predicted final size is then $n_{exp} = \tau \times N$

3. Estimate p through $p = n_{rep}/n_{exp}$

4. Repeat steps 2 and 3 above, each time using the new value of p obtained until

$$|p_n - p_{n-1}| < \epsilon \text{ and } |\beta_n - \beta_{n-1}| < \epsilon \tag{4.41}$$

   where β_n is the MLE of β given p_{n-1} and $\epsilon > 0$ is a tolerance level for the difference between successive iterations

The iterative scheme above is a combination of two point estimates methods (the MLEs in this case) using alternately the approximate likelihood (4.10) and the binomial distribution approximation in (4.29).

We apply the 4 steps above to the data described before in Table 4.1 where $N = 600$, $p = 0.5$ and $\beta = 0.0033$ with $n_{rep} = 241$ reportings. The results obtained are shown in Table 4.22 with $\epsilon = 10^{-8}$.

Contact rate β converges to the value 0.003262 which is very close to the true parameter used in the simulation of the data (0.0033). The converged value of p is 0.5038 with a true value of $p = 0.5$, confirming that our iterative scheme converges to the true parameter values of β and p.

For these data, the convergence happens after 26 iterations. It is interesting to consider how we can speed up convergence. One possibility would be to use early

Iteration	p	β	$R_0 = \beta N$	τ	n_{exp}	new p
1	1	0.00217	1.333	0.4544	272.6369	0.884
2	0.884	0.00227	1.395	0.5069	304.1293	0.792
3	0.792	0.00239	1.463	0.5581	334.8796	0.72
4	0.72	0.00250	1.537	0.6057	363.461	0.663
5	0.663	0.00262	1.610	0.6474	388.434	0.620
6	0.620	0.00274	1.682	0.6829	409.7276	0.588
7	0.588	0.00285	1.748	0.7119	427.1279	0.564
8	0.564	0.00294	1.806	0.7346	440.7321	0.547
9	0.547	0.00302	1.853	0.7517	451.0472	0.534
10	0.534	0.00308	1.893	0.7648	458.903	0.525
11	0.525	0.00313	1.923	0.7745	464.6759	0.519
12	0.519	0.00317	1.946	0.7814	468.8549	0.514
13	0.514	0.00319	1.962	0.7861	471.7078	0.511
14	0.511	0.00322	1.974	0.7896	473.7626	0.509
15	0.509	0.00323	1.982	0.7918	475.0982	0.507
16	0.507	0.00324	1.988	0.7935	476.1075	0.506
17	0.506	0.003246	1.992	0.7946	476.755	0.5055
18	0.5055	0.00325	1.994	0.7952	477.1235	0.505
19	0.505	0.003254	1.997	0.796	477.549	0.5046
20	0.5046	0.003255	1.998	0.7962	477.7606	0.5044
21	0.5044	0.003257	1.9986	0.7964	477.8635	0.5043
22	0.5043	0.003259	1.9999	0.7968	478.0741	0.5041
23	0.5041	0.0032595	2.0004	0.7969	478.1528	0.504
24	0.504	0.0032610	2.0014	0.7972	478.3082	0.5039
25	0.5039	0.0032613	2.0016	0.7973	478.3545	0.5038
26	0.5038	0.003262	2.00086	0.7970	478.2279	0.5039

Table 4.22: Iterative estimation of β and p

observation data for estimation of R_0. In the early stages of the epidemic the cumulative number of infections (infectives plus removals) grows exponentially at rate βN. Insofar as a proportion p of these are observed, the number of observed infections also grows exponentially at rate βN. Hence regardless of assumptions about p, even if it is assumed $p = 1$, and since N is assumed large, the parameter β can be reasonably estimated from the early stages of the epidemic. At least some inference about p can then be made from a comparison of the observed final size of the epidemic compared with the predicted total final size (observed plus unobserved infections) based on the already estimated value of β.

The converged values are only point estimates and there is a need for a measure of uncertainty. These can be obtained by considering the Fisher information for β and p at the converged values. But these would only be conditional measure of uncertainties.

Arguably, a better way to estimate β and p and relevant measures of uncertainty is to adopt a Bayesian approach close to that developed in Section 4.5.2.

4.8 Discussion

In this chapter, we have considered the SIR epidemic model with constant probability of reporting. Leaving aside the natural approach of full Bayesian methodology, implemented through RJMCMC, we make use of approximations to derive an approximate likelihood. Based on this approximate likelihood we have been able to propose 3 different methods to estimate the model parameters. Each of the methods gives a very good solution to the parameter estimation problem, as demonstrated with different datasets considered for application. Further confirmation came from simulation studies where different parameter values were used for data simulation. The approximation methods were also compared to the RJMCMC method and they turn out to agree very well. One advantage with the approximate methods is the time required for the algorithms to obtain converged chains. Run on the same machine, the code with approximate methods converged faster than RJMCMC. For the data with $N = 600$, the approximate methods took less than 4 hours to obtain converged chains, while for the RJMCMC it took more than 36 hours. In real-time epidemics, efficient and fast methods are preferable and the approximate methods can provide very useful tools.

Chapter 5

Varying probability of reporting

5.1 Introduction

The study in the previous two chapters assumes constant probability of reporting. However, the reporting probability for epidemics is mostly non-constant (Fraser *et al.*, 2009; Dorigatti *et al.*, 2012). Many factors are responsible for this issue:

1. Epidemiological factors: reporting of infections can be related to the severity of the disease. The rate of reporting is more likely to increase with the degree of severity of disease symptoms. Closely related factors to the severity that influence the reporting are the morbidity and mortality rates associated to the disease. The number of reported cases can also be responsible for modifications in the reporting process since when a lot of cases are known, the rate of reporting is likely to increase.

2. Socio-economic factors: the reporting of illness of individuals could also be influenced by the population connectivity or network they belong to. For example if neighbours of susceptible individuals are confirmed cases of the infectious epidemic, once some symptoms are observed, the likelihood of reporting is higher. An important social factor is media exposure. In outbreaks where media coverage is extensive, there is a higher chance of case reporting. Another factor not less negligible is the economic or financial implications of pointing out infections. This factor can be envisaged for example in cases where farmers would be constrained to close their farms or even when an individual facing significant financial loss from other activities would prefer not to report.

In this chapter, we study models that incorporate some of the factors that influence the reporting process. This is done by considering again the Markovian SIR epidemic or general stochastic epidemic and incorporating the reporting process in addition. In the following, we describe the models in turn, each with a specific reporting factor incorporated. We then provide inferential tools with applications to data to further state some conclusions and discussion related to these problems.

5.2 Models with different reporting scenarios

The physical progression of the epidemic remains as in the Markovian SIR epidemic. We assume throughout this chapter that the reporting happens at time of removal for each infected individual. We study 3 different cases of the reporting process as we describe in the following subsections.

5.2.1 Probability of reporting as a function of time

The reporting process in an epidemic is very likely to be time-dependent. As mentioned in Section 5.1, factors that could influence the reporting to change with time are the media coverage and the disease morbidity or mortality.

One approach is to assume that the probability of reporting is a step function of time. Suppose that there exist n_c change points (n_c being integer) in the reporting process and let us denote by $\boldsymbol{a} = (a_1, \ldots, a_{n_c})$ the vector of times corresponding to the n_c change points in increasing order with a_0 being the kick-off time of the epidemic and a_{n_c+1} being the end of the observation. Therefore, the reporting probability at time t is determined by

$$p(t) = \sum_{l=0}^{n_c} p_l \mathbf{1}_{[a_l, a_{l+1})}(t) \tag{5.1}$$

where p_l is the constant reporting probability in the interval $[a_l, a_{l+1})$ with $l = 0, 1, \ldots, n_c$. $\mathbf{1}_{[a_l, a_{l+1})}(t)$ is the indicator function giving 1 if $t \in [a_l, a_{l+1})$ and 0 other-

wise. The model likelihood function in this case is

$$L(\beta,\gamma,p,n_c,\boldsymbol{a};\boldsymbol{s}_{-w},s_w,\boldsymbol{r}) \propto \left\{\prod_{i\in\mathcal{I}_{-w}}\beta I(s_i^-)\right\}\exp\left(-\int_0^T \beta S(t)I(t)dt\right)$$
$$\prod_{i\in\mathcal{R}}\gamma\exp\left(-\gamma(r_i-s_i)\right)\prod_{i\in\mathcal{I}\cap\bar{\mathcal{R}}}\exp\left(-\gamma(T-s_i)\right) \quad (5.2)$$
$$\prod_{l=0}^{n_c} p_l^{t_l}(1-p_l)^{m_l-t_l},$$

where t_l is the number of reported removals in the interval $[a_l, a_{l+1})$ and m_l the total number of removals in $[a_l, a_{l+1})$ with $l = 1, \ldots, n_c$.

5.2.2 Probability of reporting as a function of the number of reported cases

We still consider the Markovian SIR epidemic, but with a different factor influencing the reporting process.

We now assume that the reporting probability depends on the number of reported cases observed. Given that the population becomes aware of the impact of the disease on a certain number of people, the reporting probability is likely to change. In principle, it is likely to increase first when many people have been reported and then probably decrease when the impact of the disease has weakened. We assume that there is n_c change points for the reporting probability and denote by $\boldsymbol{N}^r = (N_1^r, N_2^r, \ldots, N_{n_c}^r)$, the vector of numbers of reported cases at points of which the reporting probability changes. The reporting probability is assumed to be constant from one change point to the next. The number of reported cases is time-dependent and let us denote by $Rep(t)$ the number of reported cases by time t. The reporting probability can be written as

$$p\left((Rep(t)\right) = \sum_{l=0}^{n_c} p_l \mathbf{1}_{\left\{N_l^r, N_l^r+1,\ldots,N_{l+1}^r-1\right\}}\left(Rep(t)\right), \quad (5.3)$$

where p_l is the constant probability for the number of reported cases in the set $\left\{N_l^r, N_l^r+1, \ldots, N_{l+1}^r-1\right\}$ and $l = 0, 1, \ldots, n_c$. The indicator function $\mathbf{1}_{\left\{N_l^r, N_l^r+1,\ldots,N_{l+1}^r-1\right\}}(Rep(t))$ gives 1 if $Rep(t) \in \left\{N_l^r, N_l^r+1, \ldots, N_{l+1}^r-1\right\}$ and 0 otherwise.

Interestingly, the function $Rep(.)$ is an increasing function of time and we know the time at which each reporting has occurred. It is clear then that in terms of inference, the Bayesian approaches for this model and the model described in the previous subsection are similar. Indeed, each of the N_i^r values corresponds to a time a_i of change point for the reporting probability. Therefore, converting the vector of change points $\boldsymbol{N}^r$ into $\boldsymbol{a}$ as in previous subsection, we obtain again the same likelihood as in (5.2) from which a Bayesian approach can be applied for inference.

5.2.3 The probability of reporting depends on the source of infection

The reporting of infection can also be influenced by the social network an individual belongs to. One important issue that needs attention when studying epidemics is to identify patterns of the evolution of the epidemic among the population. Of course when it comes to case reporting, this is greatly affected by such patterns. To incorporate these issues into the reporting process would require knowledge of the social structure of the population.

In the approach considered in this chapter, we do not define any particular social structure for the population, but we take into account the transmission network of the disease. The assumption of homogeneously mixing population still holds. We consider the immediate influence of the source of infection on the reporting process. The following reformulation of the physical progression of the epidemic will help us understand better the reporting process.

An alternative way to model the generalised stochastic epidemic (GSE) is to consider the infectious life history $(\mathcal{I}_i, \{W_{ij}; 1 \leq j \leq N\})$ of an infective, say i, where $\mathcal{I}_i$ is the length of individual i's infectious period and $W_{ij} (1 \leq j \leq N)$ are the points of time relative to individual i's infection, at which individual i makes an infectious contact with individual j. This description is made by Neal and Roberts (2005).

For the general stochastic epidemic, $\{\mathcal{I}_i; 1 \leq i \leq N\}$ are independently and identically distributed according to $\mathbb{I} \sim \text{Exp}(\gamma)$ and $\{W_{ij}; 1 \leq i, j \leq N\}$ are independently and identically distributed according to $W \sim \text{Exp}(\beta)$. The course of the epidemic, which can be used for simulation can be described as follows given $(\mathcal{I}_i, \{W_{ij}; 1 \leq j \leq N\})$ $(1 \leq i \leq N)$. Start from the initial infectives infectious at time 0. Then let s_i be

the time at which individual i becomes infected, then $r_i = s_i + \mathcal{I}_i$ denotes the time at which individual i becomes removed. If $W_{ij} < \mathcal{I}_i$, individual i makes infectious contact with individual j at time $s_i + W_{ij}$. If individual j is still susceptible at time $s_i + W_{ij}$, individual j becomes infected, otherwise nothing happens. The above process is continued until the epidemic ceases, meaning that there are no more infectives remaining in the population.

One advantage of describing the Markovian SIR epidemic in terms of this individually-based framework is that we can clearly identify in a simulation the source of infection for each infected case. Therefore the infectious contact network for a simulated epidemic is clearly known and this helps us to build in the idea of what we will refer to as *dynamic reporting*. The reporting process can now be built in as described below.

Here, the probability of reporting for an infected individual depends on whether or not the infection of the individual that has transmitted the infection was reported or not. The probability of reporting for an infected individual increases if the individual that is the source of infection has been observed as infected. This assumption is quite realistic for example in human behaviour, where people mostly go to hospital after feeling symptoms of a particular disease, if their closest contacts are known infectious cases. It is also applicable in the case of farm epidemics where reporting is more likely to happen if the closest farms have known infections. Here we assume that there are two constant probabilities of reporting p_1 and p_2 with $p_1 < p_2$. An individual case is reported with probability p_1 if the individual's source of infection has not been reported. Also an individual is reported with the same probability p_1 if its removal happens before the removal of the source individual. It is then realistic to consider that if the source of infection has been reported, the probability of reporting for a new case increases to p_2 ($p_2 > p_1$). We also assume that the initial infected individual reports with probability p_1 since their source of infection comes from outside the population and the reporting of individuals from outside the population is not taken into account.

Considering such a reporting process with the underlying epidemic assumed to follow a Markovian SIR structure, we can obtain the likelihood of the model. We

denote by $\mathcal{N}$ the infectious contact network of the model. The likelihood is

$$\begin{aligned} L\left(\beta,\gamma,p,\mathcal{N},\boldsymbol{s}_{-w},s_w,\boldsymbol{r}\right) \propto & \prod_{i\in\mathcal{I}_{-w}} \beta I(s_i^-)\exp\left(-\int_0^T \beta S(t)I(t)dt\right) \\ & \prod_{i\in\mathcal{R}} \gamma\exp\left(-\gamma(r_i-s_i)\right) \prod_{i\in\mathcal{I}\cap\bar{\mathcal{R}}} \exp\left(-\gamma(T-s_i)\right) \quad (5.4) \\ & p_1^{n_{p_1}}(1-p_1)^{m_{p_1}} p_2^{n_{p_2}}(1-p_2)^{m_{p_2}}, \end{aligned}$$

where n_{p_i} and m_{p_i} $(i=1,2)$ are respectively the number of removal times observed (with probability p_i) and unobserved (with probability $(1-p_i)$). The need to construct the contact network when making inference will have implications on the algorithm for updating parameters. In fact, the algorithm will require more attention since every new configuration of event times will correspond to a new contact network and we will return to this issue later.

5.3 Inference

The physical progression of the model is the same as in Chapter 3. Therefore the updating of the times is kept identical to the RJMCMC algorithms described in 3.4.2. Also the full conditional distributions of the parameters β and γ are the same as in Equations (3.17) and (3.18) respectively. It remains to update the reporting probabilities in the different scenarios considered.

5.3.1 Reporting probability as a function of time

We need to know an estimate of the time-change points, to be able to estimate the reporting probabilities during the corresponding time intervals. If the change points are unknown, reversible jump MCMC methods can be applied, as was considered in a different context by Boys and Giles (2007). For simplicity, and because of the limited information in the data, we will assume here that the change points are known. This assumption is realistic in cases where, for example, information from external sources are available. Media coverage can for instance spread news about the disease at a particular time and according to how urgent the situation is, the time of news will approximately correspond to the change point in the reporting. Another possible

scenario is if the mortality rate is known to have increased considerably at a particular point in time when looking at the daily updates about the disease and there is awareness of this fact in the population.

Thus, as in the case of constant probability of reporting, we consider a beta prior for each of the probabilities p_l. Given the priors

$$p_l \sim \mathcal{B}(\alpha_{p_l}, \tau_{p_l}) \quad l = 0, 1, \ldots, n_c \tag{5.5}$$

we obtain the conditional posterior distributions

$$p_l | n_c, \boldsymbol{r}, \boldsymbol{s}_{-w}, s_w, \beta, \gamma \sim \mathcal{B}\left(\alpha_{p_l} + t_l, \tau_{p_l} + m_l - t_l\right) \quad l = 0, 1, \ldots, n_c \quad . \tag{5.6}$$

We can then simply update the reporting probabilities using Gibbs sampling.

The inference steps here also apply to the model described in Subsection 5.2.2, which, regarding inference, is equivalent to the model in 5.2.1.

5.3.2 Update of the reporting probabilities in the case of dynamic reporting

Again, with conjugate beta priors $\mathcal{B}(\alpha_{p_i}, \tau_{p_i})$, we obtain a $\mathcal{B}(\alpha_{p_i} + n_{p_i}, \tau_{p_i} + m_{p_i})$ posterior for p_i $(i = 1, 2)$. However in this case we need to estimate the infectious contact network. With each iteration of the MCMC algorithm, given the proposed event times, we associate a possible infectious contact network $\mathcal{N}$ which enables us to identify n_{p_i} and $m_{p_i} (i = 1, 2)$ in the likelihood. To associate an infectious contact network with a proposed set of times we proceed as follows in each MCMC iteration:

- sort all times in increasing order together with the corresponding individuals;
- the first infected individual has been infected from outside the population;
- the second case has been infected by the first infected individual;
- for the remaining ordered infections, assume that in the ordered set of event times and corresponding individuals, we are at infection time s_v for individual v. Possible individuals who could have infected v are those that are infected before v and are removed after time s_v;

- choose at random one individual, say u, that infects the current individual v.

Having built the network as above, we then identify the number of individual infections that have been observed with probability p_1 or p_2. We then accept this network if the corresponding times are accepted. Notice that the updating of the network is paired with the updating of the event times. In other words, to a proposed time update corresponds an infectious contact network and they are both accepted or not.

5.4 Applications

We consider the simulated dataset introduced in Section 3.5, where the physical progression parameters of the epidemic are $\beta = 0.003$ and $\gamma = 0.1$ with $n = 93$ infections ultimately happening with perfect reporting. We now simulate the reporting part of the process assuming that the reporting probability changes with time.

5.4.1 Reporting as a function of time

The reporting part is simulated by assuming a particular case where there exists $n_c = 1$ change point which happens at time $a_1 = 37.0$ (in days). We assume that the probability of reporting which is $p_0 = 0.4$ before the change, becomes $p_1 = 0.8$ after time $a_1 = 37.0$. This gives a dataset of $n_{rep} = 55$ removal times, 23 of which have been reported before a_1, and 32 after that time. The choice of $a_1 = 37$ days is motivated by the need to have a representative number of reported individuals in the two time intervals in order to be able to estimate the reporting probabilities p_0 and p_1. For the choices of p_0 and p_1, we expect a considerable increase of reporting especially if changes happen as result of extensive media coverage, jump in mortality rate or a contact having been reported. From a purely experimental reason, p_0 and p_1 are chosen far from each other so that they can be identified. If the two probabilities are too close to each other, it would be difficult to know whether the model is able to distinguish different reporting probabilities. We are therefore interested in the posterior distribution of the difference between the two probabilities and this is plotted in Figure 5.5.

Considering the data with $n_{rep} = 55$ removal times we apply the RJMCMC method

Non-informative priors $\mathcal{U}(0,1)$					
	mean	sd	2.5%	50%	97.5%
β	0.00342	0.00077	0.00210	0.00335	0.00509
γ	0.117	0.0331	0.0699	0.110	0.1980
p_0	0.428	0.105	0.272	0.413	0.688
p_1	0.878	0.099	0.634	0.902	0.995
t_0	55.689	8.641	34.000	57.000	67.000
t_1	35.991	4.280	32.000	35.000	47.000
n	91.681	7.555	72.000	94.000	100.000
R_0	3.088	1.070	1.601	2.886	5.741
$\mathcal{B}(18,27)$ for p_0 and $\mathcal{B}(40,10)$ for p_1					
	mean	sd	2.5%	50%	97.5%
β	0.00347	0.000737	0.00225	0.00340	0.00512
γ	0.112	0.0303	0.0676	0.1067	0.1856
p_0	0.405	0.054	0.305	0.403	0.516
p_1	0.824	0.0485	0.721	0.827	0.909
t_0	56.475	5.652	44.000	57.000	65.000
t_1	37.424	2.929	33.000	37.000	44.000
n	93.899	5.089	82.000	95.000	100.000
R_0	3.264	1.055	1.833	3.063	5.856
Known probabilities $p_0 = 0.4, p_1 = 0.8$					
	mean	sd	2.5%	50%	97.5%
β	0.00350	0.00070	0.00232	0.00344	0.00507
γ	0.1099	0.0284	0.0674	0.1055	0.1782
t_0	56.544	4.698	46.000	57.000	64.000
t_1	38.176	2.561	34.000	38.000	44.000
n	93.72	4.327	84.000	95.000	100.000
R_0	3.326	1.003	1.943	3.143	5.740

Table 5.1: Posterior estimates in the case of complete epidemic assuming step function for the reporting probability and using RJMCMC.

described in Subsection 3.4.2 for updating the times with the corresponding reporting probability updated as in 5.3.1. In fact, a Metropolis-Hastings within Gibbs algorithm is implemented as follows:

- Update β and γ following Gibbs steps using Equations (3.17) and (3.18);
- Update Event times following RJMCMC algorithm described in Subsection 3.4.2;
- For each accepted event times, count the number of removed individuals before and after the change point a_1;
- Identify the number of reported and unreported cases before and after a_1;

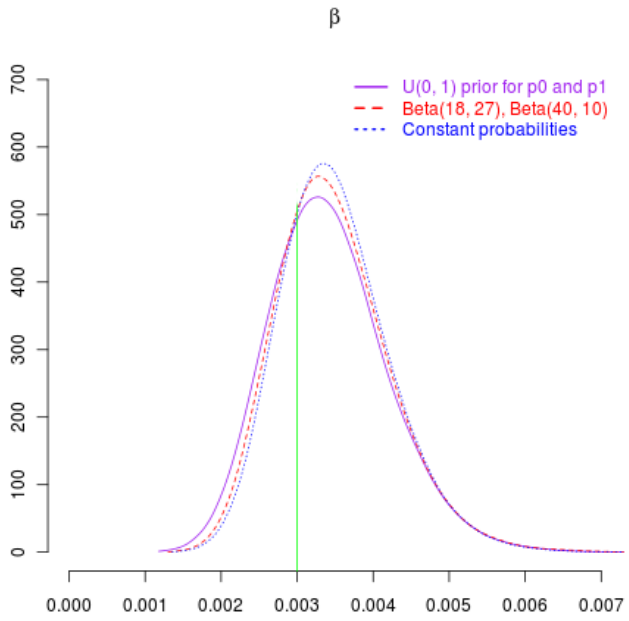


Figure 5.1: Posterior density of β when using RJMCMC and different prior distributions for p_0 and p_1: $\mathcal{U}(0,1)$ for the two probabilities (purple solid line); Beta distributions with mean 0.4 and variance 0.0052 for p_0, and mean 0.8 and variance 0.00313 for p_1 (red dashed line); constant reporting probabilities (blue dotted line).

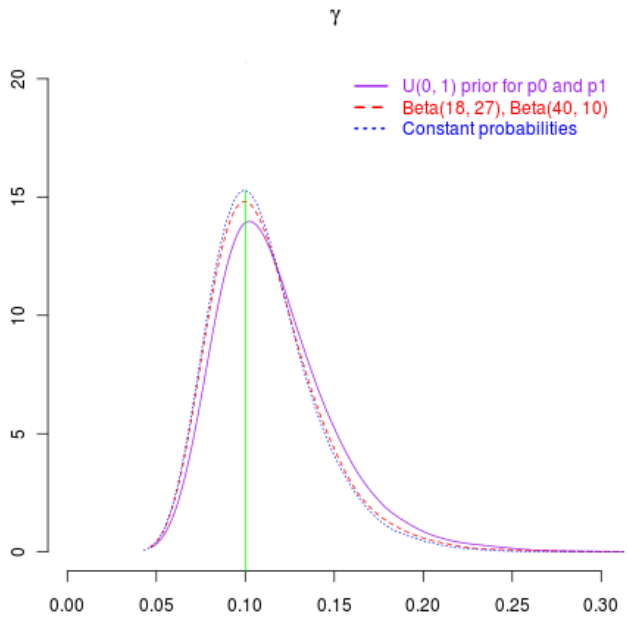


Figure 5.2: Posterior density of γ when using RJMCMC and different prior distributions for p_0 and p_1: $\mathcal{U}(0,1)$ for the two probabilities (purple solid line); Beta distributions with mean 0.4 and variance 0.0052 for p_0, and mean 0.8 and variance 0.00313 for p_1 (red dashed line); known reporting probabilities (blue dotted line).

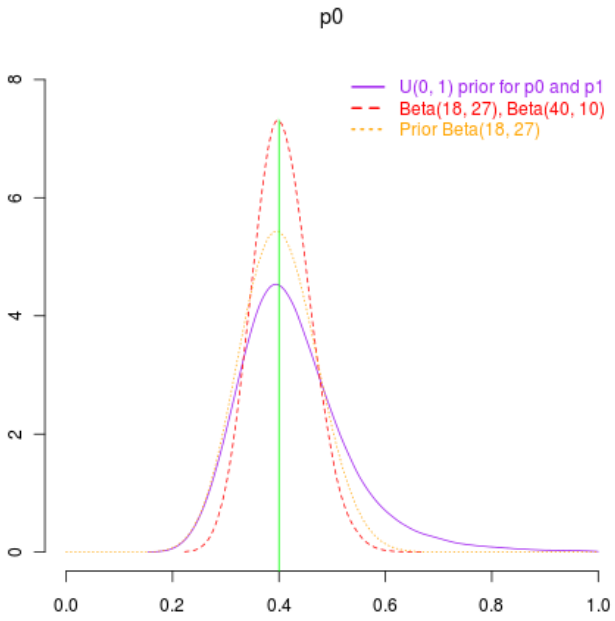


Figure 5.3: Posterior density of p_0 when using RJMCMC and different prior distributions for p_0 and p_1: $\mathcal{U}(0,1)$ for the two probabilities (purple solid line); Beta distributions with mean 0.4 and variance 0.0052 for p_0, and mean 0.8 and variance 0.00313 for p_1 (red dashed line).

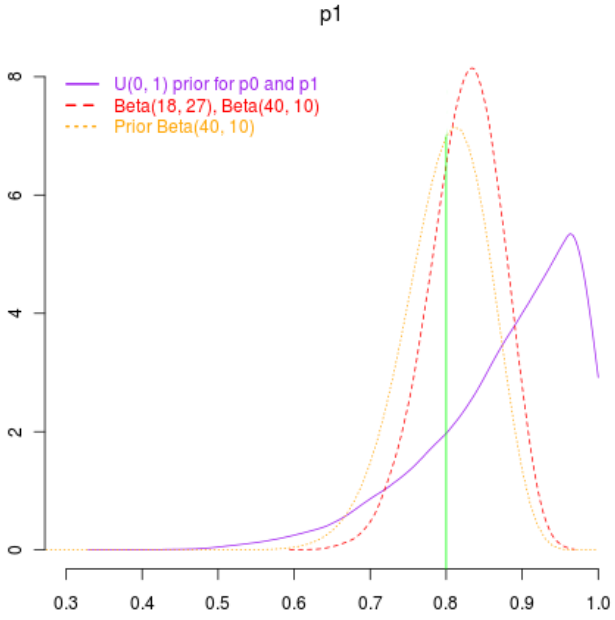


Figure 5.4: Posterior density of p_1 when using RJMCMC and different prior distributions for p_0 and p_1: $\mathcal{U}(0,1)$ for the two probabilities (purple solid line), Beta distributions with mean 0.4 and variance 0.0052 for p_0 and mean 0.8 and variance 0.00313 for p_1 (red dashed line).

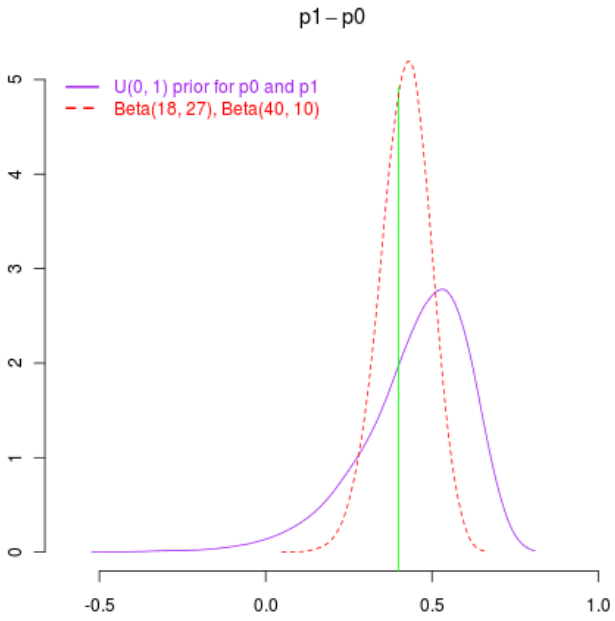
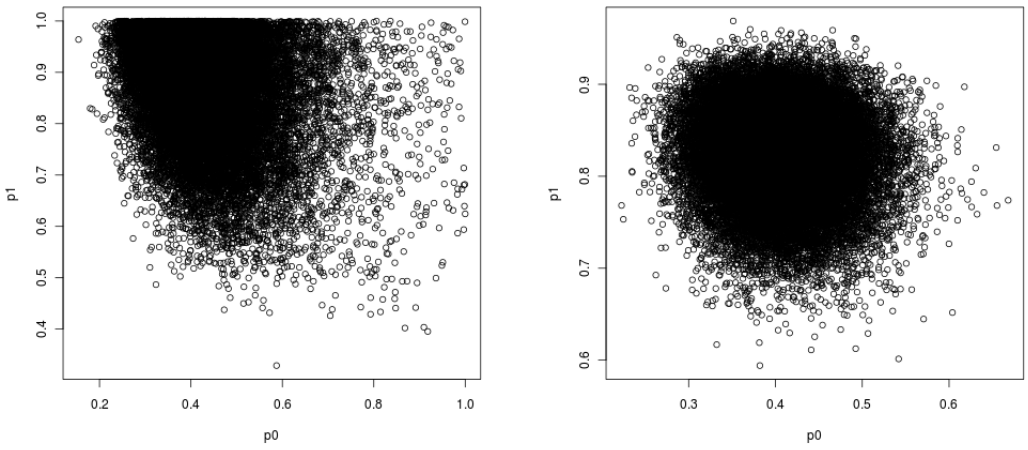


Figure 5.5: Posterior density of the difference $p_1 - p_0$ when using RJMCMC and different prior distributions for p_0 and p_1: $\mathcal{U}(0,1)$ for the two probabilities (purple solid line), Beta distributions with mean 0.4 and variance 0.0052 for p_0 and mean 0.8 and variance 0.00313 for p_1 (red dashed line).



(a) (b)

Figure 5.6: Bivariate plot of the posterior distributions of p_0 and p_1 (p_0 v p_1) when using $\mathcal{U}(0,1)$ priors on p_0 and p_1 ((a)) and Beta distributions with mean 0.4 and variance 0.0052 for p_0 and mean 0.8 and variance 0.00313 for p_1 ((b))

- Update the reporting probabilities following Equations (5.6);
- Repeat the above steps until convergence.

The above algorithm is applied to the data and the posterior distributions of the parameters are summarised in Table 5.1. We also include in Table 5.1 results from prior sensitivity analysis on the reporting probabilities. The posterior densities are plotted in Figures 5.1-5.4.

The means of the parameters β and γ are slightly bigger than the true parameter values $\beta = 0.003$ and $\gamma = 0.1$, causing a slight increase in the reproduction number R_0. However, all these estimates agree with the true parameter values since the latter are all well within the credible intervals. The mean values of the posterior estimates of β and γ are not very much influenced by the different priors used. However we can notice a decrease in the standard deviations when knowledge of the reporting probabilities becomes more accurate, as visible from Table 5.1 and the plots of the posterior densities. Regarding the estimation of the reporting probability, the mean of the posterior estimate of p_1 looks considerably bigger than the true parameter value. The posterior distribution of p_1 is left-skewed with a higher mode than the true parameter value when using non-informative priors, confirming that there is high uncertainty related to the estimation in this case. As expected, p_0 and p_1 are more accurately estimated with informative prior distributions on each of them. Indeed, with the priors $\mathcal{B}(18, 27)$ for p_0 and $\mathcal{B}(40, 10)$ for p_1, the means are centered on the true parameter values $p_0 = 0.4$ and $p_1 = 0.8$, with variances respectively 0.0052 and 0.0031. The variances of the posterior distributions obtained are respectively 0.0029 and 0.00235 for p_0 and p_1. It is clear that the posterior estimates of p_0 and p_1 provide much narrower distributions than the priors, meaning that in the case of informative priors, there is still a lot to gain from the posterior estimates. These can also be noticed from the plot of the density of the difference $p_1 - p_0$ in Figure 5.5, where the variance of the difference is smaller in the case of more informative priors on p_0 and p_1. The bivariate plot of the two probabilities p_0 and p_1 in Figure 5.6 is another illustration where the posterior samples are more clustered when the priors are more informative (5.6(b)) compared to non-informative priors (5.6(a)). The posterior density of the difference $p_1 - p_0$ in Figure 5.5 also shows that the algorithm is able to distinguish between the two probabilities, particularly when there is some prior knowledge.

5.4.2 Dynamic reporting

Estimation of the model parameters

Applications on the model with dynamic reporting are also considered using simulated data. The simulation of the data is based on the algorithm described in Subsection 5.2.3, where we are able to track the source of infection for each infected individual. The physical progression of the epidemic is still as in the Markovian SIR system with the transition probabilities given in Equations (3.12) and (3.13). The contact and removal rates used for the simulation of the data are $\beta = 0.003$ and $\gamma = 0.1$ respectively. We obtain a total number of infections and removals $n = 93$, after $T = 90$ days. As described in the model (Subsection 5.2.3), two reporting probabilities are considered. If the case acting as the source of infection of an individual has not been reported, such individual's removal time is reported with probability $p_1 = 0.5$ and not reported with probability $1 - p_1$. For an individual whose source of infection is known to have been reported, the reporting probability becomes higher and we set it to be $p_1 = 0.8$. With such probabilities specified, the total number of reported infections with probability p_1 is $n_{p_1} = 31$ and $n_{p_2} = 23$ with probability p_2, giving a total of $n_{rep} = 54$ reported removal times. The parameters of the model with the different sizes are summarised in Table 5.2.

N	β	γ	R_0	n	p_1	p_2	n_{p_1}	n_{p_2}
100	0.003	0.1	2.97	93	0.5	0.8	31	23

Table 5.2: True parameters for data simulation and different sizes obtain in the case reporting depends on the source of infection.

We recall that due to under-reporting, the simulated data consist of the 54 removal times reported. We then aim to infer about the rates β and γ, the reporting probabilities p_1 and p_2 and the construction of the chain informing about the source of infections as detailed in Subsection 5.3.2.

The priors for β and γ are chosen to be non-informative: $\beta \sim \text{Ga}(0.001, 0.001)$ and $\gamma \sim \text{Ga}(0.001, 0.001)$. We assume different prior distributions for the reporting probabilities in order to study the sensitivity of the posterior estimates to prior choice.

We start by assuming a uniform prior on $p_1 \sim \mathcal{U}(0,1)$, but assume a beta prior for $p_2 \sim \mathcal{B}(8,2)$. With a uniform distribution on p_2 ($\mathcal{U}(0,1)$), it was difficult to obtain a non-degenerate MCMC chain, for reasons we will point out later in Section 5.5. More informative priors for both reporting probabilities p_1 and p_2 have been considered. We assume a $\mathcal{B}(5,5)$ prior for p_1 (mean = 0.5, variance = 0.0227) and a $\mathcal{B}(12,3)$ prior for p_2 (mean = 0.8, variance = 0.01). We also assume a $\mathcal{B}(10,10)$ prior for p_1, giving a mean of 0.5 and variance of 0.012, for p_1 and a $\mathcal{B}(24,6)$, where the mean is 0.8 and the variance is 0.0052 for p_2. The extreme case of known reporting probabilities are also considered and the results are summarised in Table 5.3. The posterior densities of the model parameters are plotted in Figures 5.7-5.11.

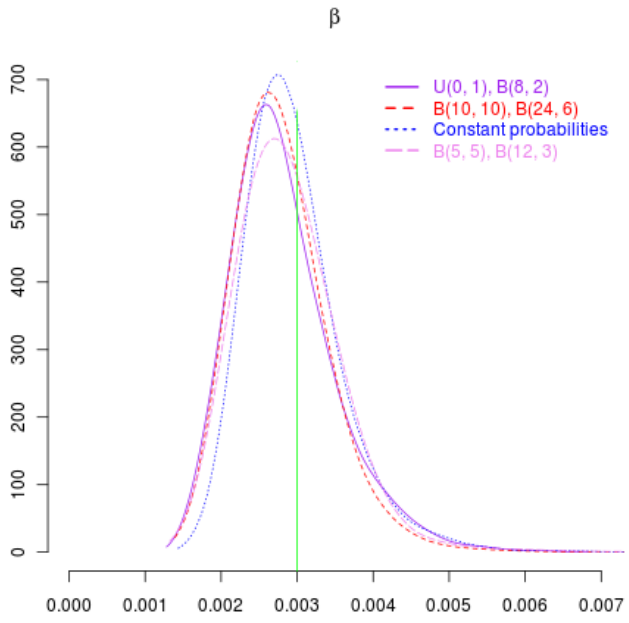


Figure 5.7: Posterior density of β when using RJMCMC and different prior distributions for (p_1, p_2): $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ (purple solid line); $(\mathcal{B}(5,5), \mathcal{B}(12,3))$ (violet dashed line); $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (red dashed line); and fixed known reporting probabilities (blue dotted line).

The means of the posterior distributions of β and γ are close to the true parameter values; while their shape are skewed to the right. When using non-informative prior for the reporting probability p_1, its posterior mean ($\bar{p}_1 = 0.57$) is higher than the true

$\mathcal{U}(0,1)$ for p_1 and $\mathcal{B}(8,2)$ for p_2					
	mean	*sd*	2.5%	50%	97.5%
β	0.00282	0.00069	0.00175	0.00271	0.00448
γ	0.099	0.036	0.050	0.092	0.188
p_1	0.577	0.171	0.331	0.536	0.963
p_2	0.816	0.118	0.531	0.839	0.975
n_{p_1}	31.72	5.45	22.00	31.00	43.00
n_{p_2}	22.28	5.45	11.00	23.00	32.00
n	84.839	12.68	59.000	88.00	100.00
R_0	3.19	1.60	1.54	2.76	7.939
$\mathcal{B}(5,5)$ for p_1 and $\mathcal{B}(12,3)$ for p_1					
	mean	*sd*	2.5%	50%	97.5%
β	0.00288	0.00068	0.00185	0.00280	0.00434
γ	0.098	0.030	0.0597	0.0923	0.179
p_1	0.507	0.108	0.322	0.497	0.749
p_2	0.823	0.093	0.605	0.837	0.963
n_{p_1}	31.12	5.18	19.00	31.00	41.00
n_{p_2}	22.87	5.18	13.00	23.00	35.00
n	89.83	9.87	67.00	93.00	100.00
R_0	3.17	1.36	1.60	2.80	6.50
$\mathcal{B}(10,10)$ for p_1 and $\mathcal{B}(24,6)$ for p_1					
	mean	*sd*	2.5%	50%	97.5%
β	0.00279	0.00063	0.00184	0.00272	0.0042
γ	0.097	0.028	0.0589	0.0918	0.165
p_1	0.512	0.082	0.367	0.506	0.691
p_2	0.808	0.068	0.659	0.814	0.922
n_{p_1}	31.71	4.65	23.00	32.00	41.00
n_{p_2}	22.29	2.73	13.00	22.00	31.00
n	89.62	8.70	70.00	91.00	100.00
R_0	3.07	1.24	1.68	3.76	6.27
Known probabilities $p_1 = 0.5,\ p_2 = 0.8$					
	mean	*sd*	2.5%	50%	97.5%
β	0.00295	0.00064	0.0020	0.00286	0.00452
γ	0.099	0.0285	0.0627	0.094	0.173
n_{p_1}	32.49	4.18	24.00	33.00	40.00
n_{p_2}	21.516	4.18	14.00	21.00	30.00
n	92.60	6.415	78.00	94.00	100.000
R_0	3.14	1.18	1.77	2.84	6.33

Table 5.3: Posterior estimates of the model parameters in the case of complete epidemic and assuming that the reporting probability depends on the source of infection.

value ($p_1 = 0.5$). This is due to the right-tail in the distribution of p_1 as we can see in the density plot in Figure 5.9. The right-tail in the distribution of p_1, with the large variance associated reflect the high uncertainty related to the estimation here, as we discuss in the next section. In all cases, we notice that the true parameter values are

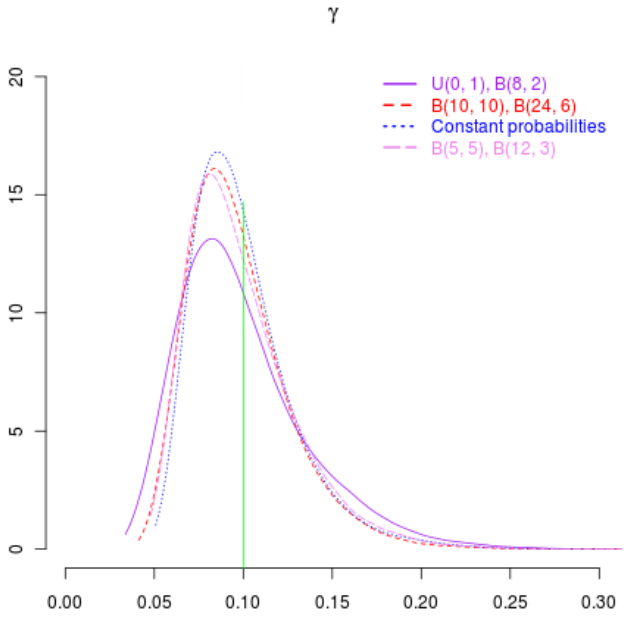


Figure 5.8: Posterior density of γ when using RJMCMC and different prior distributions for (p_1, p_2): $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ (purple solid line); $(\mathcal{B}(5,5), \mathcal{B}(12,3))$ (violet dashed line); $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (red dashed line); and fixed known reporting probabilities (blue dotted line).

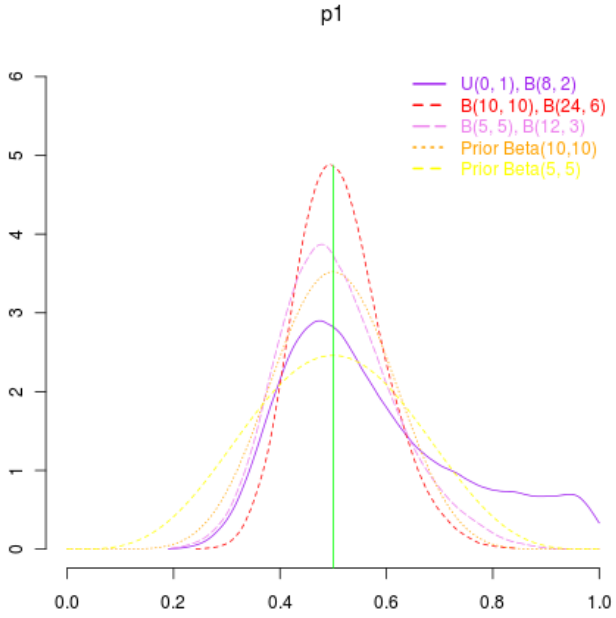


Figure 5.9: Posterior density of p_1 when using RJMCMC and different prior distributions for the couple (p_1, p_2): $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ (purple solid line); $(\mathcal{B}(5,5), \mathcal{B}(12,3))$ (violet dashed line); and $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (red dashed line).

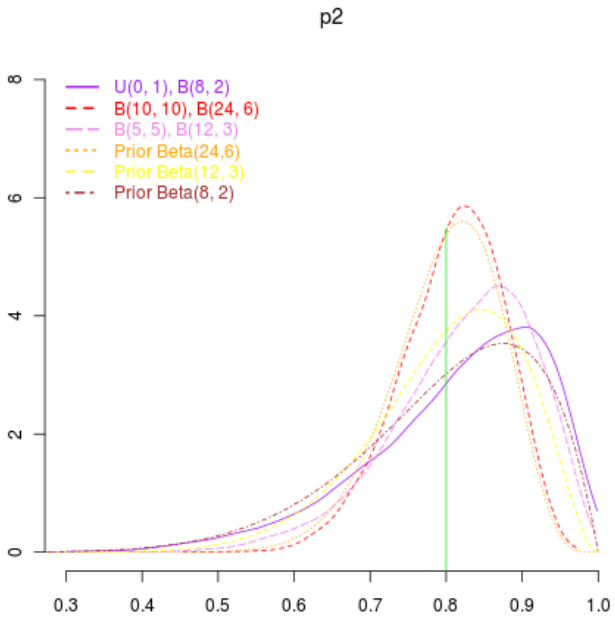


Figure 5.10: Posterior density of p_2 when using RJMCMC for different prior distributions for the couple (p_1, p_2): $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ (purple solid line); $(\mathcal{B}(5,5), \mathcal{B}(12,3))$ (violet dashed line); and $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (red dashed line).

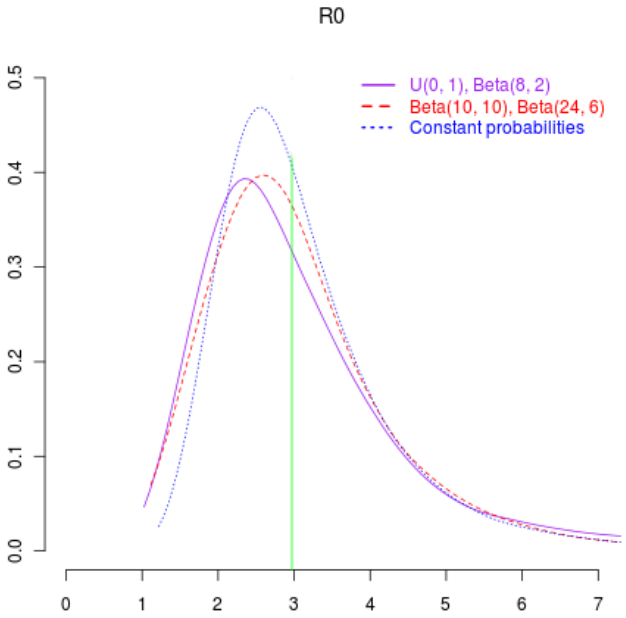


Figure 5.11: Posterior density of R_0 when using RJMCMC and different prior distributions for the couple (p_1, p_2): $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ (purple solid line); $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ (red dashed line); and fixed known reporting probabilities (blue dotted line).

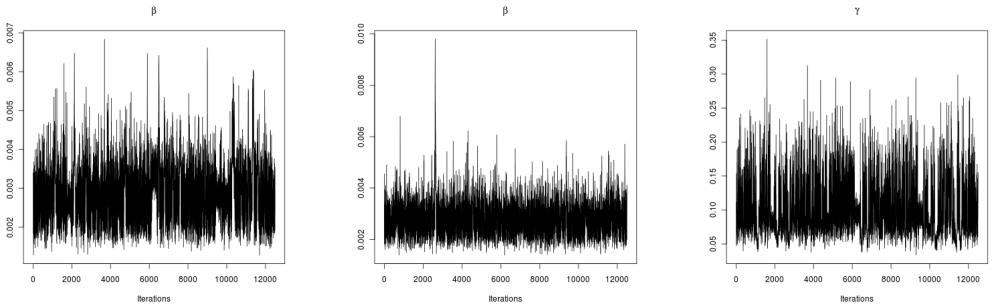
well contained in the credible interval of the posterior distributions of each parameter. We compare the prior and posterior variances of the reporting probabilities. For a $\mathcal{U}(0,1)$ prior on p_1, the posterior variance of p_1 is 0.0292, which is much smaller than the prior variance (0.0833). By assuming a $\mathcal{B}(8,2)$ distribution for p_2, the mean of the prior distribution on p_2 is 0.8 and the variance is 0.0145. From the posterior distribution, the mean of p_2 is 0.816 and the variance is 0.0139. Despite the high uncertainty related to the estimation, we still gain information from the posterior distributions, even though we need some prior knowledge on the reporting probabilities. In Table 5.4.2, we compare the prior and posterior variances of the reporting probabilities p_1 and p_2 when more informative priors are used. the posterior variances are smaller in all the cases than the prior ones.

$(\mathcal{B}(5,5), \mathcal{B}(12,3))$ for (p_1, p_2)			$(\mathcal{B}(10,10), \mathcal{B}(24,6))$ for (p_1, p_2)		
	Post. variance	Prior variance		Post. variance	Prior variance
p_1	0.0117	0.0227	p_1	0.0067	0.0119
p_2	0.0086	0.01	p_2	0.0046	0.0516

Table 5.4: Posterior and prior variances of the reporting probabilities in cases of different priors

The cases of informative and known reporting probabilities provide smaller posterior variances for the parameters β and γ. Information in the prior distributions for the reporting probabilities is automatically linked to knowledge about the number of reported individuals whose source of infection have reported or not, i.e. individuals who have been reported with probability p_1 or p_2. This is reflected in the posterior estimates of the number of individuals with reporting probability p_1 (n_{p_1}) and similarly with reporting probability p_2 (n_{p_2}). Indeed, the variances of the distributions of n_{p_1} and n_{p_2} decrease with more information on p_1 and p_2.

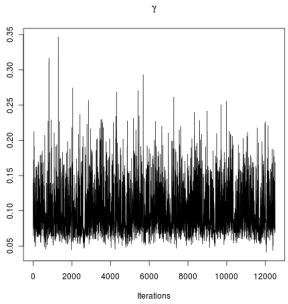
The convergence of the Markov chains are assessed by looking at the chain traces. In Figure 5.12, we plot the sample traces of the chains. The chains mix with more difficulty when the priors on p_1 and p_2 are non-informative because of the limited information here.



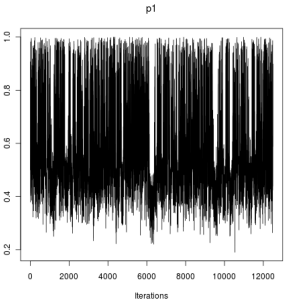
(a) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$

(b) $(\mathcal{B}(10,10), \mathcal{B}(24,6))$

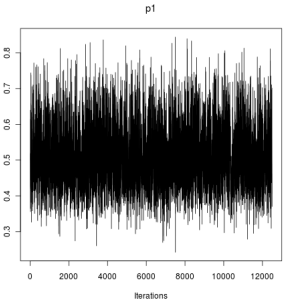
(c) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$



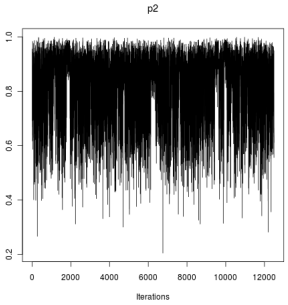
(d) $(\mathcal{B}(10,10), \mathcal{B}(24,6))$



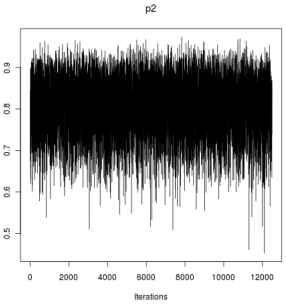
(e) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$



(f) $(\mathcal{B}(10,10), \mathcal{B}(24,6))$



(g) $(\mathcal{U}(0,1), \mathcal{B}(8,2))$



(h) $(\mathcal{B}(10,10), \mathcal{B}(24,6))$

Figure 5.12: Sample traces for β, γ, p_1 and p_2 after burn-in period of 1000 iterations and a thinning of 20 samples, in the case of completed epidemic with reporting depending on the source of infection and using $(\mathcal{U}(0,1), \mathcal{B}(8,2))$ ((a)), (c), (e) and (g)) and $(\mathcal{B}(10,10), \mathcal{B}(24,6))$ ((b), (d), (f) and (h) for (p_1, p_2).

Estimation of the infectious contact network

Another interesting aspect of the estimation is the reconstruction of the origin of infection for the infected individuals in the MCMC iterations, since each time update can result to a new possible source of infection to be considered. As explained in Subsection 5.3.2, proposal of new infection times in the RJMCMC, implies a new proposal for the infectious contact network. In order to make inference about the source of infection for each individual, there is a need to label the outside population individual responsible for the first infection and the origin of infection for individuals that have

not been estimated to be infected in each RJMCMC iteration. Let us denote by "O" the outside individual that infected the first individual in the population. Individual "O" can, for instance, represent the bacteria responsible for the first infection. If an individual in the population is susceptible in the chain of estimation, we create a pseudo source of infection for such individual that we denote "S". It is important to know that there is no meaning for source of infection for a susceptible individual and that this is created purely to keep track of the state of individuals in terms of source of infection. The part of the algorithm that updates the contact network is described as follows:

- At a given iteration of the MCMC algorithm, we propose a set of event times;
- All the times are sorted in the increasing order with their corresponding individuals
- Start building the infectious contact network by considering that the individual with the earliest time of infection has been infected from outside the population, therefore individual "O";
- The second earliest infection can only have been infected by the first infected individual since the population is closed;
- Any given infected individual can only have been infected by an individual that is still infectious and have been infected before the current individual: select one of the possible infectors at random;
- Move through all the infection times and attach a source of infection to all the infected individuals;
- Affect "S" as source of infection to all the individuals that are susceptible at this iteration;
- Identify the reported individuals;
- If the source of infection of an individual has not been reported, the individual in question has reported with probability p_1;

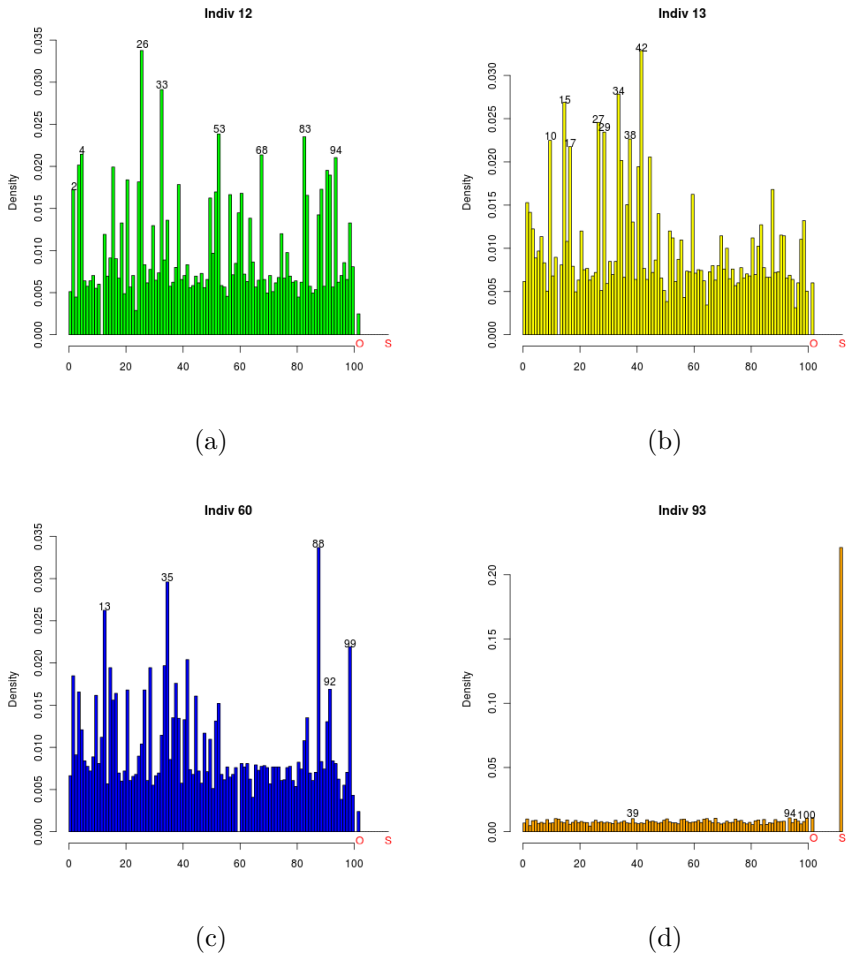
- If the source of infection of an individual has been reported but the individual got removed before its source of infection, such individual has reported with probability p_1;
- Any other reported individual has been reported with probability p_2;
- The proposed event times and the contact network are accepted or rejected together.
- The steps above are repeated every time the set of event times has been changed with new proposition in the RJMCMC;

When running the RJMCMC algorithm, we record the infectious network at each iteration and have for each individual a distribution of his source of infection. To illustrate this, we choose 4 individuals ($12, 13, 60$ and 93) and comments on the inference made on each of them. Individuals 12, 13 and 60 were reported as infected while 93 was not reported infected. This means that we have no information about individual 93 regarding his state and he will be treated at the beginning of the RJMCMC as susceptible.

We plot in Figure 5.13 the histogram of the possible infectors of the individuals 12, 13, 60 and 93 and compute the posterior probabilities of their first five respective infectors in Tables 5.5-5.8. Since the "true" origins are known for each individual from data simulation, we compare them with the posterior distribution that estimates the source of infection.

Individual 12		
	Source of infection	Posterior probability
1^{st}	26	0.0337
2^{nd}	33	0.0291
3^{rd}	53	0.0238
4^{th}	83	0.0235
5^{th}	68	0.0214

Table 5.5: The estimated first five source of infection for individual 12 with their corresponding posterior probabilities



(a) (b)

(c) (d)

Figure 5.13: Histograms of the estimation of the source of infection for individuals 12 ((a)), 13 ((b)), 60 ((c)) and 93 ((d)).

Individual 13		
	Source of infection	Posterior probability
1^{st}	42	0.0329
2^{nd}	34	0.0278
3^{rd}	15	0.0268
4^{th}	27	0.0245
5^{th}	29	0.0234

Table 5.6: The estimated first five source of infection for individual 13 with their corresponding posterior probabilities

We can see from the histogram in Figure 5.13(a) and Table 5.5 that the source of infection for individual 12 is more likely to come from individuals 26 or 33. Individual 12 is truly infected by individual 68 who turns out to be the fifth possible source of infection of individual 12 in the inference. We recall that we are working with simulated datasets and therefore we have the data from perfect reporting with the

Individual 60		
	Source of infection	Posterior probability
1^{st}	88	0.0337
2^{nd}	35	0.0296
3^{rd}	13	0.0262
4^{th}	99	0.0219
5^{th}	42	0.0204

Table 5.7: The estimated first five source of infection for individual 60 with their corresponding posterior probabilities

Individual 93		
	Source of infection	Posterior probability
1^{st}	S	0.2212
2^{nd}	O	0.0110
3^{rd}	94	0.0108
4^{th}	100	0.0106
5^{th}	39	0.0102

Table 5.8: The estimated first five source of infection for individual 93 with their corresponding posterior probabilities

true source of infections that we are using to make the comparisons here. Figure 5.13(b) shows the histogram of the source of infection of individual 13. The estimation suggests that this individual is more likely to have been infected by individuals 42 or 34 while in reality, he was infected by individual 17. Individual 17 turns out to be eighth highest possible source in the estimation. The true source of infection of individual 60 which is individual 92 turns out to be fourteenth in the ranking from the estimation. Individual 93 who was not reported infected is more likely to be inferred as non-infected from the estimation since the estimation suggests that most likely, he remains susceptible. Overall, we can see that individual 93 is considered non-infected with probability 0.2212, while different infection sources are associated with it with smaller probabilities. This is to be expected since when this individual is considered infected in the estimation, there are many possible sources of infection that can be associated with it. If individual 93 is infected, its source of infection is likely to be from outside the epidemic (individual O) or individual 94 while the real source of infection is individual 78.

From the results above of the estimated source of infection we do not always recover the simulated source of infection for each individual. This can be due to the random choice of the source of infections in the algorithm as mixing of the population

in the model is homogeneous. Also, all the individuals have the same infectivity and susceptibility depending on their state. More discussion of this model is given in the following section.

5.5 Discussion

In this chapter we have studied 3 models, 2 of which are equivalent as regards to inference. The models have a common underlying infection and removal processes which is the Markovian SIR model. With the same physical progression, the models differ in the corresponding reporting processes. In the case of time-dependent model (Subsection 5.2.1) and number-of-reported-cases model (Subsection 5.2.2), the reporting processes were argued to coincide in term of inference in Subsection 5.2.2. Therefore, assuming that we know the point in time at which the reporting changes, we showed that we can make good inference about the physical progression parameters β and γ, and also the reporting probabilities. The prior sensitivity analysis showed that even if we have some prior information on the reporting probabilities, we still gain a lot when we make inference by obtaining more accurate results with narrower variances.

The model with reporting process depending on the source of infection turned out to be the one which results in more uncertainty about the estimations. Indeed, as explained in Subsection 5.3.2, a new proposal of infection times corresponds to a new proposal of infectious contact network, where both event times and network are updated together (or not). This affects the mixing of the Markov chain as the acceptance probability in the MCMC is subject to the proposed times and the resulting proposed network. It is possible that the new times proposed agree with the data but the contact network built with it does not allow acceptance of the proposed times. The contact networks are built by uniformly selecting the possible individuals, according to the proposed times, that can be the source of infection for each individual. It therefore provides a large number of possibilities each time that the network is built and when some of these possibilities are not accepted, the chain does not move from its current state.

The model is homogeneous and therefore all individuals are assumed to have the same infectivity and susceptibility. With a model where some individuals are known to have higher infectivity for instance, the infectious contact network could be inferred

taking into account their high probability of being a possible source of infection for many individuals, giving then a more informed building of the contact network, which in turn will give better mixing of the chain and estimation. Moreover, the population is homogeneously mixing. Therefore, there is no a priori information on possible paths of transmission of the disease. With more structure on the mixing of the population, the construction of the infectious contact network will be made using appropriate probability for associating a source of infection to each individual case.

Chapter 6

Introduction to the spatial aspect: Under-reporting on $\mathbb{Z}$

6.1 Introduction

The aim of this chapter is to study the effect of under-reporting on $\mathbb{Z}$ as an introduction to epidemics on graphs with reporting processes. A large variety of epidemic models on graphs exist going from theoretical results (Kuulasmaa, 1982; Kuulasmaa and Zachary, 1984; Kuulasmaa and Mollison, 1985) to statistical inference (Britton and O'Neill, 2002; Caimo and Friel, 2011). Some other models deal with the mixture of space and time as these two aspects can be part of real epidemic processes (Gibson *et al.*, 2006; Filipe *et al.*, 2009; Cook *et al.*, 2007). In these models, statistical analyses are conducted assuming perfect reporting i.e. the data reflect all the (possible) infections that have happened throughout the time observation of the epidemic. Our goal here in this chapter is to study the possibility of under-reporting, its influence on inference and to provide methodologies to account for it, making more reliable inference.

We therefore define the Markovian SIR epidemic on the line $\mathbb{Z}$ as detailed in the following section (6.2). Like all the models described in the previous chapters for study in this thesis, the model here also will consist of two main parts: physical progression of the epidemic and reporting process. The modelling will take into account whether or not the event (infection or removal) times are known or not. After incorporating the reporting process in the physical progression of the epidemic the likelihood of the

model will be derived for the two cases of known or unknown event times. Bayesian inference will be considered using either final size data or temporal data to obtain posterior distributions for the model parameters. A comparison will be made between perfect reporting cases and under-reporting cases with simulated data. Further comparisons will be between the two cases of known and unknown event times to draw useful conclusions based on the inference results.

6.2 Model

6.2.1 General description

The model is still the Markovian SIR but this time considered on graph. We consider the line $\mathbb{Z}$ where each coordinate represents a site susceptible to a disease. Each site can be thought of as a plant on a line in a field. It can be assumed that plants in a field are more often in a form of lattice. In this model here, the assumption is that we are not considering vertical links between vertices on a lattice. Each vertex can only transmit disease to its neighbour(s) that is on the same horizontal line. Therefore, in terms of disease transmission, the horizontal lines are left independent of each other and it suffices to study the epidemic on a single line. This assumption is realistic because many plant diseases spread in rows, as the distance between plants in rows is much smaller than between columns. We assume now that one of the sites becomes infected from an infectious disease and is at the same time likely to spread it.

We assume that an infective plant emits germs following a Poisson process with rate β. In other words the probability that an infective individual emits an infectious germ in a small interval of time dt is given by

$$\Pr(\text{ a germ is emitted in } (t, t+dt)) = \beta dt + o(dt) \tag{6.1}$$

An infected site becomes removed following a Poisson process with rate γ, meaning that the length of the infectious period for each infective site is exponentially distributed with parameter γ. We assume independence between the Poisson processes.

For such a model, each individual can only infect its neighbours i.e the spread is to the nearest neighbours only. But apart from the first infective site which is likely

to infect its two neighbours, any other infective in the population can only infects its neighbour which is still susceptible. We also assume that no removed individual is subject to further infection and that no reinfection is possible in this model.

We assume for the reporting process that each infected site is reported with the same constant probability p. The under-reporting makes the model very exciting to study since the physical progression on its own seems straightforward. For inference purpose, we are going to consider two scenarios consisting of known or unknown event times data in order to complete the specification of this model.

6.2.2 All the infection and removal times are unknown

First insight

The first analysis we consider for this model is to assume that we do not have any time information available. We assume that during the course of the epidemic, none of the event times is observed. Hence in the case of perfect reporting, we only know that the site was infected without a record of when the infection happened and when the removal occurred. With the reporting process defined above, if $p < 1$, we can only observe some of the infected sites. By denoting n_r and n_l respectively the right and left end points reported, we obviously deduce that all the sites or individuals in between were infected during the epidemic. We can then make a crude estimation of the reporting probability. But one main concern in this model is the possible unreported infections that happened beyond n_r and n_l. To make further analysis on estimating the model parameters we write down the likelihood.

Likelihood

Let us assume that the first infected site is 0 and it was able to infect its two neighbours. Actually, in our model, we can imagine the 2 Poisson processes as being 3 independent Poisson processes where the first with rate $\beta/2$ can emit germs to infect the right neighbour, the second with the same rate infects the left neighbour, and the third with rate γ of getting removed. We can then compute probabilities of removal and infections.

The probability that the initial infected individual infects both neighbours is

$$p_{0,-1,1} = \left(\frac{\beta}{\beta+\gamma}\right)\left(\frac{\beta}{\beta+2\gamma}\right) \tag{6.2}$$

where the first term takes into account the fact that the germ has been emitted and will surely infects one neighbour and the second term explains the probability of infection of the other neighbour. From here we are left with independent events of spread on the left and right. Each of the infections either on the left or right until we reach the end points n_r and n_l happens with probability

$$p_{1,2} = \frac{\beta}{\beta+2\gamma}. \tag{6.3}$$

We now need to take into account individuals that could have been infected and unreported in the likelihood. Each of them is infected with probability $p_{1,2}$ and unreported with probability $1-p$. So considering that there are m of them ($m \in \mathbb{N}$) on the left we deduce that the probability of unreported m infected cases is

$$p_{m,u} = \left(\frac{\beta}{\beta+2\gamma}(1-p)\right)^m \left(\frac{\gamma}{\gamma+\beta/2}\right) \tag{6.4}$$

where the last term expresses that the epidemic has ceased progressing to one side. Because m can take any integer value, the contribution from the left unreported infected individuals is the sum over $\mathbb{N}$ of the probability $p_{m,u}$. It is easy to compute since there is a geometric series when summing over $\mathbb{N}$. Some algebra gives

$$p_{l,u} = \sum_{m\in\mathbb{N}} p_{m,u} = \frac{2\gamma}{2\gamma+\beta p}. \tag{6.5}$$

$p_{l,u}$ is the probability of unreported infections that happen to one side. One interpretation of this result in (6.5) that we will call "correction factor" in the likelihood is the following. $p = 1$ corresponds to a case of perfect reporting and since we have observed the left infected end point on the line for instance , we know that its removal occurred before having to emit germs. This probability of getting removed before the possibility of emitting germ is $2\gamma/(2\gamma+\beta)$ which is $p_{l,u}$ where $p = 1$. The reasoning is the same for the right unobserved infected sites.

Combining all the different contributions lead to the likelihood

$$L(\beta,\gamma,p)=\left(\frac{2\gamma}{2\gamma+\beta p}\right)^2\left(\frac{\beta}{\beta+\gamma}\right)\left(\frac{\beta}{\beta+2\gamma}\right)^{n_r-n_l-1}p^{n_{rep}}(1-p)^{n_{unrep}} \tag{6.6}$$

where n_{rep} and n_{unrep} are respectively the number of reported and unreported infections between n_l and n_r with n_l and n_r included, giving

$$n_{rep}+n_{unrep}=n_r-n_l+1. \tag{6.7}$$

Again as in previous chapters, the term $p^{n_{rep}}(1-p)^{n_{unrep}}$ is the information from the reporting process between n_l and n_r. The likelihood in Equation (6.6) is derived using one example of how the epidemic could spread with the reporting process associated. But in fact, it does not matter whether or not we know if the epidemic moves from the first infected site to the right and left. All that matters is that we know the reported sites and more importantly the right and left end-points n_r and n_l respectively, and the likelihood turns out to be the same as in Equation (6.6).

6.2.3 Infection and removal times of reported infected sites known

The second case we consider in the study of this model is to assume that we know the infection and removal times of the reported infected individuals. In the model above, once an individual infects its neighbour it plays no further role in the spread of the epidemic if there is no other susceptible to infect. Such individual remains in the state of infectious (I) until he gets into the state of removal (R) after an exponentially distributed length of time. Furthermore, we assume that no reinfection is possible, meaning that in the case all the neighbours of an infected individual are already infected, he just remain infectious until being removed without infecting again the neighbours. We therefore assume that when an individual infects its neighbour, it stays infectious at its site until it dies out meaning that its removal time corresponds to the time it is unable to emit germs even though emitting a germ does not influence the spread of the disease having already infected its neighbour. The advantage of this assumption is to have an idea of how long germs are emitted before ceasing. Using the same notations as before, let us include some other ones to write down the likelihood

with the times.

The processes are the same as described above and we denote by
$\boldsymbol{s} = (s_{n_l}, s_{n_l+1}, \ldots, s_0, s_1, \ldots, s_{n_r})$ the vector of infection times of the respective sites
$\mathcal{Z} = \{n_l, n_l+1, \ldots, 0, 1, \ldots, n_r\}$ with the corresponding removal times
$\boldsymbol{r} = (r_{n_l}, r_{n_l+1}, \ldots, r_0, r_1, \ldots, r_{n_r})$. The model likelihood can be written as:

$$\begin{aligned} L(\beta, \gamma, p; \boldsymbol{s}, \boldsymbol{r}) &= 2\beta' \exp\{-2\beta'(s_1 - s_0)\}\, \beta' \exp\{-\beta'(s_{-1} - s_1)\} \\ &\quad \prod_{i=-1}^{n_l+1} \beta' \exp\{-\beta'(s_{i-1} - s_i)\} \prod_{i=2}^{n_r} \beta' \exp\{-\beta'(s_i - s_{i-1})\} \\ &\quad \prod_{i \in \mathcal{Z}} \gamma \exp\{-\gamma(r_i - s_i)\} \\ &\quad \left(\frac{2\gamma}{2\gamma + \beta p}\right)^2 p^{n_{rep}} (1-p)^{n_{unrep}} \end{aligned} \tag{6.8}$$

where $\beta' = \beta/2$. Equation (6.8) is an augmented likelihood since some of the sites have not been reported as infected and therefore we do not know their infection and removal times. The first expression of this likelihood takes into account the fact that the first infected site is able to infect its both neighbours. The second and third expressions are simply the product of infections and removals probabilities coming from different sites. The last expression in this likelihood contains the correction factor in Equation (6.5). It is used so that we can avoid to impute times from events beyond n_r and n_l. The last term in such last expression is just the information coming from the reporting process between n_l and n_r. A simplified version of likelihood (6.8) can be written as

$$\begin{aligned} L(\beta, \gamma, p; \boldsymbol{s}, \boldsymbol{r}) &= 2\beta'^{n_r + |n_l|} \exp\{-\beta'[(s_{n_r} - s_0) + (s_{n_l} - s_0))]\} \\ &\quad \gamma^{n_r + |n_l| + 1} \exp\left\{-\gamma \sum_{i \in \mathcal{Z}} (r_i - s_i)\right\} \\ &\quad \left(\frac{2\gamma}{2\gamma + \beta p}\right)^2 p^{n_{rep}} (1-p)^{n_{unrep}} \end{aligned} \tag{6.9}$$

after some very easy algebra.

6.2.4 Relationship between likelihoods

One main question of interest is how the two likelihoods in Equations (6.6) and (6.9) are related. The two likelihoods should be the same if we transform the likelihood with time (6.9) to obtain a likelihood without time. In other words, integrating the likelihood in Equation (6.9) over times should give the likelihood in Equation (6.6). In this Subsection, we describe briefly how it is possible to show that (6.9) leads to (6.6) and provide important constraints on the event times of the spread of the disease that are useful for efficient inference methodology as detailed in Subsection 6.3.2.

To be able to do the integration, we need to identify that there are some constraints on the times as follow: for any site i,

$$i \geq 1, \quad s_i < s_{i+1} < r_i < \infty \tag{6.10}$$

and for any site j,

$$j \leq -1, \quad s_i < s_{i-1} < r_i < \infty. \tag{6.11}$$

Assuming that the first infection caused by site 0 goes to the right, we have

$$s_0 < s_1 < s_{-1} < r_0. \tag{6.12}$$

Under these constraints, the integration needs to be done in a systematic manner. By integrating over first r_i and then s_i starting from the both end sites, we end up with the expression in Equation (6.6). For example, the integration from the right starts with respect to r_{n_r} such that $s_{n_r} < r_{n_r} < \infty$, then goes with respect to s_{n_r} such that $s_{n_r-1} < s_{n_r} < r_{n_r-1}$ and then with respect to r_{n_r-1} and so on. These integrations tell us that it is actually possible to write the likelihood in Equation (6.9) without the unknown times, particularly the times of the sites that have not been reported as infected. In other words, the likelihood (6.9) can be transformed and written as a function of the times of reported sites only. But this requires an identification of the unreported sites with their times and the integrations would not be easy to handle. The approach here is general to be applied to any data in a formal way.

6.3 Inference

The method of inference depends on the available data, whether the event times are known or not. But in both cases, we adopt the Bayesian methodology by defining a prior distribution on the model parameters and obtain their posterior distributions through MCMC algorithm. It is important to point out that the estimation here does not require a use of RJMCMC. Indeed, the likelihoods in Equations (6.13) and (6.9) contain a correction factor that takes into account all the possible infections that are beyond the reported sites. Without such correction factor in Equation (6.5), we would be considering to input, infections that are beyond the end-points n_r and n_l which would imply a change of dimension in the parameter space through data augmentation and therefore a use of trans-dimensional MCMC. The correction factor is very useful, helping to use simple MCMC, hence providing a computationally fast estimation method for the parameters of interest.

6.3.1 Case of unknown event times

The data here consists of reported infected sites and for inference, we are only interested in the end-points n_r and n_l with the number of reported and unreported infections in between that are n_{rep} and n_{unrep}. This implies that the likelihood we can use to make inference is (6.6). With the information in the data, it is not possible to identify γ from β. Inference in such a situation in the case of perfect reporting for epidemics where the data consist only of final size of the epidemic has been studied in different applications (Demiris and O'Neill, 2006; Neal, 2010; Britton *et al.*, 2011) and the distribution of the infectious period with the parameters are assumed known. We need to assume γ known; $\gamma = 1$ without loss of generality. An equivalent way to look at likelihood (6.6) is that it is essentially a function of the ratio β/γ. Therefore, (6.6) can be rewritten as

$$L(R,p) = \left(\frac{2}{2+pR}\right)^2 \left(\frac{R}{R+1}\right) \left(\frac{R}{R+2}\right)^{n_r-n_l-1} p^{n_{ob}}(1-p)^{n_{unob}}. \tag{6.13}$$

where $R = \beta/\gamma$.

It simply remains to define a prior distribution on R and p and obtain their posterior distributions through MCMC.

6.3.2 Case of known times from reported sites

Inference is still done in the Bayesian framework. We define a prior on β, γ and p to sample from their posterior distributions. In this particular case here, there is a need to update the model parameters and the event times of the unreported sites between the end points n_r and n_l. For the unreported infection and removal times in $\mathcal{Z}$, we propose an infection and the corresponding removal time for each site. Both proposed times are accepted or rejected simultaneously.

Let $\mathcal{U}$ denote the set of sites whose event times have not been reported in $\mathcal{Z}$. The details of the algorithm are the following:

- Choose a site (let us say k) at random in $\mathcal{U}$.

- If $k < 0$, because of the constraint in Equation (6.11), we propose an infection time for k uniformly in (s_{k+1}, r_{k+1}). We then correspond a removal time using the fact that the removal time of this individual should be greater than the infection time of his neighbour he has infected. Therefore the proposal distribution for the removal time r_k can be chosen to be uniformly distributed in (s_{k-1}, T). In theory, here $T = \infty$ but in practice we can find an upper bound for the removal times. The acceptance probability is

$$A_- = \begin{cases} 0 & \text{if } s_k > r_k \text{ or } s_k > s_{k-1}; \\ \min\left\{1, \frac{L^{new}}{L^{old}}\right\} & \text{otherwise.} \end{cases}$$

- If $k > 0$, using the constraint in Equation (6.10), we propose an infection time for k uniformly in (s_{k-1}, r_{k-1}) and then propose a corresponding removal time uniformly in (s_{k+1}, T). Therefore the acceptance probability is

$$A_+ = \begin{cases} 0 & \text{if } s_k > r_k \text{ or } s_k > s_{k+1}; \\ \min\left\{1, \frac{L^{new}}{L^{old}}\right\} & \text{otherwise.} \end{cases}$$

Here, we can clearly identify β from γ by considering temporal data. Again, MCMC algorithm can be used to estimate the parameters β, γ and p.

6.4 Applications

6.4.1 Data

We simulate an epidemic on $\mathbb{Z}$ based on the infection process in (6.1) and an exponential lifetime for the disease. The parameters are chosen to be $\beta = 20$ and $\gamma = 1$. The set of infected and removed sites $\mathcal{I}$ in the case of perfect reporting on this simulation is all sites from -18 to 7 i.e. $\mathcal{I} = \{-18, -17, \ldots, 6, 7\}$. But we do not have perfect reporting in our study. Now, with the reporting probability $p = 0.5$, the set of reported infected sites is

$$\mathcal{O} = \{-18, -15, -13, -11, -8, -6, -5, -1, 0, 3, 4, 5, 6\}. \tag{6.14}$$

According to our notation, we have $n_r = 6$, $n_l = -18$ so that the set of sites that we know have been infected is $\mathcal{Z} = \{-18, -17, \ldots, 5, 6\}$. Therefore, the set of unreported sites between n_r and n_l is

$$\mathcal{U} = \mathcal{Z} \setminus \mathcal{O} = \{-17, -16, -14, -12, -10, -9, -7, -4, -3, -2, 1, 2\}. \tag{6.15}$$

Because we have a simulated dataset, we will also assume perfect reporting i.e. the set $\mathcal{I}$ with the event times considered when needed for comparison purpose. However, inference here consists mainly to use the reported data $\mathcal{O}$ to estimate the parameters. The results for the two cases of known and unknown event times are given with Bayesian analysis with comparison between them and also with the parameters used to simulate the data.

6.4.2 Results in the case of unknown times

As discussed in Subsection 6.3.1, in the case of unknown event times, the data consists of the reported infected sites and we can only infer about the rate $R = \beta/\gamma$.

With the data $\mathcal{O}$, we run MCMC algorithm using the likelihood in (6.13). We set $\gamma = 1$ to make inference on β which is equivalent to R and this allows us to compare results with the case, the times are known. We assume an improper prior for β ($\pi(\beta) \propto 1/\beta$) and use a random walk update to sample from the posterior distribution of β. The prior distribution of p is set to be $\mathcal{B}(\alpha_p, \nu_p)$ where for the

results below, $\alpha_p = \nu_p = 1$. The posterior distributions of β and p are summarised

	mean	sd	2.5%	50%	97.5%
β	42.274	39.562	8.243	29.572	153.590
p	0.486	0.096	0.300	0.485	0.675

Table 6.1: Posterior estimates of β and p with right reported end-point $n_r = 6$ and left one $n_l = -18$, unknown event times and $n_{rep} = 13$ reported sites out of $n_r + n_l + 1 = 25$

in Table 6.1. The reporting probability p is very well estimated as most information about p is contained in the binomial term in the likelihood. In our particular case here we have know that $n_{rep} = 13$ sites are reported as infected out of 25.

The posterior distribution of β is quite right skewed. It is actually the case if we plot the likelihood times prior. Notice that inference here can be made by classical statistics by finding the MLE of both parameters β and p. Nevertheless, the true parameter of β is very well in the credible interval. Also the variance of the distribution of β is large suggesting that there is a lot of uncertainty related to the estimation. The median provides a better point estimate for β as it is usually the case for skewed distributions. The data here only inform about the reported sites and therefore contain little information about the rate β at which germs are emitted. This allows such high uncertainty in the estimation.

6.4.3 Results in the case of known times

In this case now, we assume that our data consist of $\mathcal{O}$ and that the infection and removal times of each site in $\mathcal{O}$ are known. Again we use the same improper prior for β with the same non-informative prior for p as previously.

Case γ known

We assume $\gamma = 1$ as in the case of unknown times. This assumption is made to allow us to be able to make comparisons of the results. Table 6.2 contains the summary statistics of the posterior distributions of β and p. By looking at Table 6.2, all the true values of our parameters are very well contained in the credible intervals of the

	mean	sd	2.5%	50%	97.5%
β	24.314	5.114	15.338	23.958	35.338
p	0.487	0.097	0.299	0.486	0.676

Table 6.2: Posterior estimates of β and p with right reported end-point $n_r = 6$ and left one $n_l = -18$, known event times for the reported infected sites and $n_{rep} = 13$ reported sites out of $n_r + n_l + 1 = 25$

marginal posterior distributions. Also the mean of the distributions are very close to the true values.

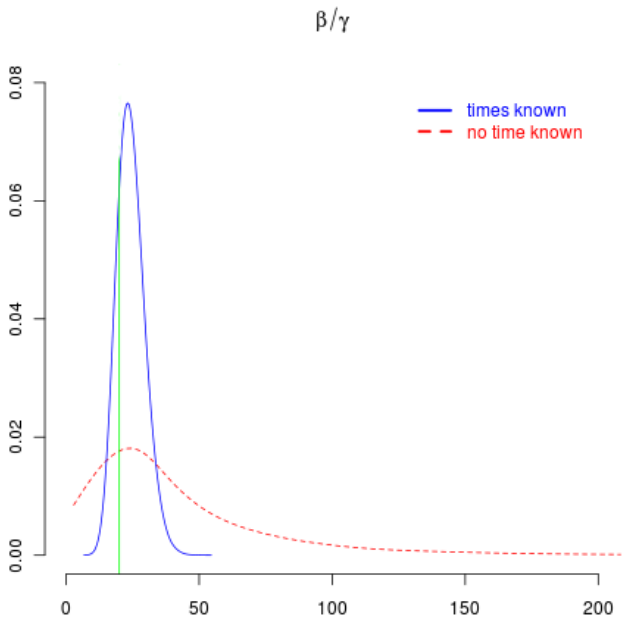


Figure 6.1: Posterior density of β in the cases of: unknown event times for the reported sites (red dashed line) and known event times for the reported sites (blue solid line)

The estimations of p in the case of known times compared to unknown times are very similar as we can see from the plots of the posterior densities in the two cases in Figure 6.2. This confirms the observation when deriving the likelihoods in the two cases that the informations about p is mostly contained in the binomial term and that the information in the correction factor about p is little. On the other, there is a huge difference in the estimations of β. It is obvious that there is a better estimation of β in the case of reported infected event times known. The skewness of

the posterior distribution of β is greatly reduced when the reported infected event times are known compared to the case of unknown times. The plots of the posterior distributions in Figure 6.1 emphasises these comments. The standard deviation is considerably reduced. Also the mean of the posterior distributions are significantly different. When none of the times is known it is far higher than the true value. The median in the case times are non-reported gives a better information about the true parameter value due to the right-skewness.

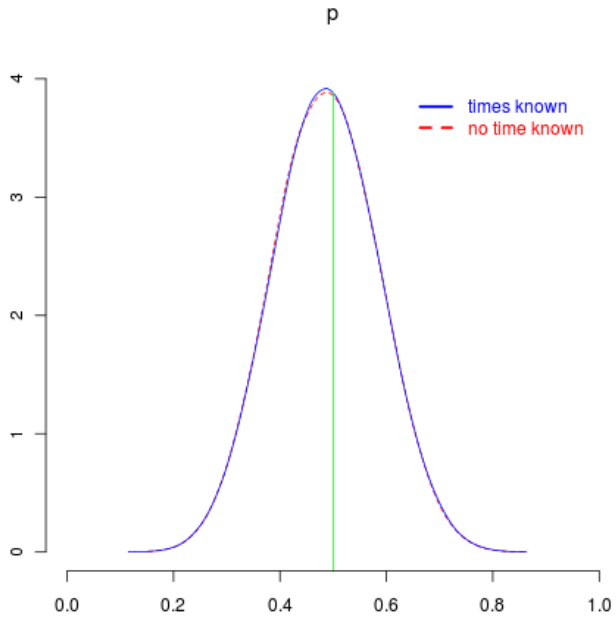


Figure 6.2: Posterior density of p in the cases of: unknown event times for the reported sites (red dashed line) and known event times for the reported sites (blue solid line)

The results here, in this statistical analysis, show that the more information there exist, the more accurate are the estimations. We should therefore encourage investments on data collection so that we have as much information as possible.

Case γ unknown

Again with our data consisting of $\mathcal{O}$ and the event times of the sites in $\mathcal{O}$ are known, we now assume that γ is unknown. The knowledge of the times ensures that β and γ are identifiable. We assume a gamma prior distribution for the removal rate γ and

choose the shape and the rate parameters of the prior such that it is completely non-informative ($\text{Ga}(0.001, 0.001)$). Keeping the same prior distributions for β ($\pi(\beta) \propto 1/\beta$) and p ($\mathcal{U}(0,1)$) as before, we obtain the posterior distributions summarised in Table 6.3. The true parameter values are all contained in the credible intervals of the

	mean	sd	2.5%	50%	97.5%
β	24.363	5.134	15.387	23.975	35.419
γ	1.204	0.259	0.758	1.183	1.769
p	0.487	0.096	0.301	0.488	0.674

Table 6.3: Posterior estimates of β, γ and p with right reported end-point $n_r = 6$ and left one $n_l = -18$, known event times for the reported infected sites, γ unknown and $n_{rep} = 13$ reported sites out of $n_r + n_l + 1 = 25$.

obtained posterior distributions. As expected, the posterior distribution of p is not influenced by the consideration of γ unknown. This is explained by the independence between the physical evolution process of the epidemic and the reporting process which simply informs about the sites that have been reported infected. Compared to

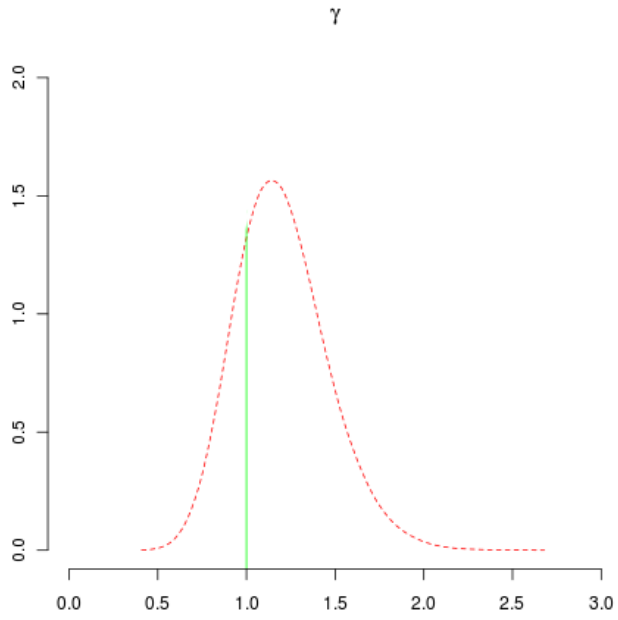


Figure 6.3: Model on $\mathbb{Z}$: posterior density of γ

the case γ is known when looking at Tables 6.2 and 6.3, the standard deviation of β is slightly wider. Such variability is expected since knowing γ reduced some uncertainty in the estimation of β.

6.4.4 Comparisons with perfect reporting

We assume perfect reporting and compare the results to provide insight of well inference can be made in this model when there is perfect reporting. The results are compared with the cases above where under-reporting is known to exist.

Case of unknown times

With the simulations made in Subsection 6.4.1, let us assume that the reporting probability was $p = 1$ and the data consist of all the observed sites that were infected $\mathcal{I}$. We fix $\gamma = 1$, the prior of β is still improper and we run the MCMC algorithm to obtain the summary statistics in Table 6.4. Looking at Tables 6.1 and 6.4, the

	mean	*sd*	2.5%	50%	97.5%
β	40.069	38.508	7.909	27.875	147.482

Table 6.4: Posterior estimate of β with right observed end-point $n_r = 7$ and left one $n_l = -18$ and perfect reporting ($p = 1$) with unknown event times

estimation of β looks better in the case of perfect reporting. The skewness to the right is reduced and so is the standard deviation. The credible interval of β is still wide due to the right skewness. But in general, there is not a big difference in the results of the two tables 6.1 and 6.4. The data used in this result is not significantly different from the case of under-reporting. Only the right end site $n_r = 7$ is not identified as infected in the case of under-reporting.

Case of known times

The data here consist of all the infected sites $\mathcal{I}$ with all their infection and removal times known. Keeping the same prior distributions we obtain the posterior distri-

	mean	sd	2.5%	50%	97.5%
β	23.954	4.895	15.394	23.625	34.416

Table 6.5: Posterior estimate of β with right observed end-point $n_r = 7$ and left one $n_l = -18$ and perfect reporting ($p = 1$) and $\gamma = 1$ with known event times

	mean	sd	2.5%	50%	97.5%
β	24.021	4.952	15.235	23.700	34.630
γ	0.990	0.189	0.658	0.978	1.394

Table 6.6: Posterior estimates of β and γ with right observed end-point $n_r = 7$ and left one $n_l = -18$ and perfect reporting ($p = 1$) with known event times

butions in Tables 6.5 in the case γ is assumed known ($\gamma = 1$) and 6.6 where γ is estimated as well.

Comparing the distributions of β in Tables 6.2 and 6.5 at once and Tables 6.3 and 6.6, we notice that β is better estimated in the case of perfect reporting. The standard deviations are smaller and the means are closer to the true parameter values. There is a better estimation of β if we know all the times compare to if only some of them are known. Basically, more informations about the times provide a more accurate estimation of the parameter β. It is obviously expected that the true standard deviation in the case $p = 1$ is less than the case $p < 1$ and the means of the distribution for β and γ in the case $p = 1$ are closer to the true parameters values than the case $p < 1$.

6.5 Discussion

The study in this chapter shows how we can define an SIR epidemic on $\mathbb{Z}$ and incorporate a reporting process in it. The statistical inference provides a considerable difference between temporal and final data. It is clearly better if the times of infection and removal of the sites are known since such information provides more accurate estimation of the rates of infection and removal. The modelling on $\mathbb{Z}$ presents the advantage of identifying infections between two reported sites; hence providing a very good idea about the reporting rate. For possible infections beyond the reported cases at the end-points reported on the left and right, we were able to associate a correction factor which enable us to avoid trans-dimensional MCMC algorithm for inference.

Chapter 7

Conclusions and Further Research

7.1 Conclusions

In this thesis, we presented methods for statistical inference for stochastic epidemic models with under-reporting. The analyses were mainly conducted in the Bayesian framework with two broad questions investigated. First we analysed possible bias that can appear when making inference in the case where under-reporting exists and the data are treated as no under-reporting was occurring. We showed that the extent to which the infection rate is under-estimated, while the influence of the reporting process on the estimation of the removal rate remains small. Secondly, we allowed under-reporting to be modelled and developed various methodologies of inference that account for the under-reporting and helped overcome the problem of under-estimation of the infection rate. The methods are flexible and can be extended to more elaborate and realistic epidemics.

In the first chapter, we presented the motivation behind modelling epidemics and the reasons that under-reporting is such an important aspect to account for in inference problems. We described the considered stochastic epidemic model as representing the physical progression of the disease, namely the Markovian stochastic SIR model. We also briefly described stochastic models that are direct extensions of the Markovian SIR model and for which it is fairly straightforward to incorporate reporting processes that would extend the statistical methodologies accounting for under-reporting.

Chapter 2 contains the literature review for statistical inference for infectious disease data. The nature of epidemic data led us to the use of Bayesian inference.

The theory of Bayesian inference is then briefly described with the MCMC algorithm that is very useful in Bayesian computation. Because of the under-reporting problem, we introduced a Bayesian inference method for missing data cases. An important part of the computational aspect was the trans-dimensionality of the MCMC method since the size of the state space was unknown due to under-reporting.

In Chapter 3, we first presented a characterisation of SIR models, putting various models in a single probabilistic framework. We considered the Markovian SIR model and added a reporting process. The removal time of each infected individual was reported with constant probability. We demonstrated the under-estimation of the infection rate when there exists under-reporting but it is not taken into account when making inference. We moved on to inference with a developed RJMCMC algorithm that allowed us to impute the unknown event times that are the result of under-reporting. This novel algorithm turned out to help remove the bias which is otherwise introduced when under-reporting exists but it is not considered.

In Chapter 4, we considered a similar model to that of Chapter 3 but this time assuming that we knew the infection times of the reported individuals. The likelihood of the model was approximated with constant probability of reporting. Three different approximate methods of inference were designed and compared with a full Bayesian analysis using RJMCMC. All the methods provided very good results and turned out to be faster than the use of RJMCMC, therefore providing tools that can be used in real-time epidemics for which this model is an appropriate representation.

The fifth chapter of this thesis explored the same physical progression of the epidemic (Markovian SIR model) but considering more realistic reporting probabilities. We therefore looked at cases of a time-dependent reporting probability and also the possibility that the reporting probability was dependent on the source of infection for each individual. We assumed a step function for the time-dependent reporting probability. By assuming that we knew the change-points of this function and using RJMCMC, we were able to make inference about the parameters of the physical progression and the reporting probabilities. It turned out that when reporting is dependent on the source of infection, there is a need to have increased *a priori* information on the reporting probabilities to help the mixing of the chains.

In Chapter 6, we studied the under-reporting problem on the $\mathbb{Z}$-axis again using

SIR modelling. The model was studied in the cases of final size data and temporal data. The statistical analysis in the Bayesian framework did not require the use of RJMCMC as we were able to find a correction factor that incorporated information from vertices that have not been reported as infected and that are beyond the reported vertices on the line $\mathbb{Z}$. Inference based on the temporal data provided more accurate results, as expected, compared to the case of final size data.

7.2 Suggestions for further research

The physical progression of the epidemic considered throughout this thesis assumed an exponential distribution for the infectious period, making the model Markovian. The methods of inference can easily be extended to cases of other non-negative distributions for the infectious period such as Weibull (Streftaris and Gibson (2004a)) and Gamma (O'Neill and Becker (2001); Jewell *et al.* (2008)).

In Chapter 4, the assumption of a constant probability of reporting was very important for the approximations (4.4), (4.8) and (4.10) to be made. In the case of a varying reporting probability, for instance as a function of time or number of reported cases, it would be interesting to explore how approximations can be made to again speed up any inference methods that can come from Bayesian data augmentation methodology. One interesting question that needs to be answered as well is the extent to which the approximations can be made if the infection times are unknown.

Departing from constant probability of reporting to more realistic reporting processes added more uncertainty to the estimations. However, the models were studied with simplified assumptions. For instance when assuming a step function for the reporting probability that is time-dependent in Subsection 5.2.1, we also assumed that the change points are known. Further work related to this model would be to explore the extent to which we can let the model estimate what the change points are, and therefore make inference about the number of different reporting probabilities. The methodology to apply can be motivated from work by Boys and Giles (2007) where in a multitype SEIR model, the removal rate was assumed to be a step function of time and inference was made about the number of change points with the removal rate in each time interval. A more general change-point detection problem is discussed by Adams and Mackay (2005).

It would be interesting to explore the idea behind the source-of-infection model on different scenarios where, for instance, the infectivity varies (Streftaris and Gibson, 2004b), there exist different types of severity which can lead to varying infectivity (Ball and Britton, 2005), or the susceptibility varies (with respect to individuals O'Neill and Becker (2001), or time Gibson *et al.* (1999, 2004)). In these models, the weight for choosing a possible source of infection for a given individual will not be uniform between the candidates, as the more infectious individual would be more likely to be the infector.

On a broader picture, there is a question of whether or not we should be departing from the assumption of constant probability of reporting given specific data. Different reasons related to socio-economic and epidemiological factors can motivate the consideration of defining more complex reporting processes rather than a constant one. There is a need to explore such possibility from competing models, to decide which one fits "best" the data. Model assessment and selection tools are therefore needed. This, in fact, is a more general topic that needs more progress in the Bayesian framework.

Extensions to more complex graphs like tree or lattice need to be considered in future research on models of epidemics on graphs with reporting process incorporated. There exist attempts in the literature on inference for models on lattices (Bailey *et al.*, 2000; Sander *et al.*, 2003), but not with explicit modelling of reporting process. In terms of structure, as we studied for $\mathbb{Z}$, trees also present, in the case of under-reporting, the advantage of knowing the source of infection for all reported cases as two vertices are connected by exactly one simple path. However, deeper study is required to make a formal description of the model and state clear conclusions after inference. The model on the lattice $\mathbb{Z}^2$ appears much more complex since between two reported vertices, there are several possible paths that infectious germs can take to move from one vertex to the other. Experiments are probably required first to provide an insight about how conclusions can be drawn. Other ideas would be first to put restrictions on the graphs considering for instance directed graphs. Further investigations are needed in the future towards this direction.

References

R. P. Adams and D.J.C. Mackay. Bayesian online change-point detection. Technical report, Technical Report at University of Cambridge, 2005.

C. L. Addy, I. M. Longini, and M. Haber. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47(3):961–974, 1991.

M. Aitken. Posterior Bayes Factors (with discussion). *J. Roy. Statist. Soc. B*, 53: 111–142, 1991.

R. Albert and A.-L. Barabàsi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer, 2000.

D. J. Bailey, W. Otten, and C. A. Gilligan. Saprotrophic invasion by the soil-borne fungal plant pathogen Rhizoctonia solani and percolation thresholds. *New Phytologist*, 146(3):535–544, 2000.

N. T. J. Bailey, editor. *The Mathematical Theory of Infectious Diseases and its Applications*. 2nd ed. London: Griffin, 1996.

N. T. J. Bailey and A. S. Thomas. The estimation of parameters from population data on the general stochastic epidemic. *Theoretical Pop. Biol.*, 2:253–270, 1971.

F. Ball and T. Britton. An epidemic model with exposure-dependent severities. *J. Appl. Probab.*, 42(4):932–949, 2005.

F. G. Ball. The threshold behaviour of epidemic models. *J. Appl. Probab.*, 20:227–241, 1983.

F. G. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Probab.*, 18:289–310, 1986.

F. G. Ball and O. D. Lyne. Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.*, 177–178:333–354, 2002.

F. G. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7:46–89, 1997.

N. G. Becker. *Analysis of infectious disease data. Monographs on Statistics and Applied Probability.* Chapman & Hall, London, 1989.

N. G. Becker and P. Yip. Analysis of variation in an infection rate. *Australian Journal of Statistics*, 31:42–52, 1989.

N. G. Becker, T. Britton, and P. D. ONeill. Estimating vaccine effects on transmission of infection from household outbreak data. *Biometrics*, 3:467–475, 2003.

K. Bennett, J. Phillipson, P. Lowe, and N. Ward. The impact of Foot and Mouth crisis on rural firms: A survey of microbusinesses in the North East of England. *Research Report University of Newcastle*, 2001.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, New York, 1994.

O. N. Bjornstad, B. F. Finkenstadt, and B. T. Grenfell. Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184, 2002.

A. Bouma, I. Claassen, K. Natih, D. Klinkenberg, C. A. Donnelly, G. Koch, and M. van Boven. Estimation of transmission parameters of H5N1 avian influenza virus in chickens. *PLoS Pathog*, 5, 2009.

R.J. Boys and P.R. Giles. Bayesian inference for SEIR epidemic models with time-inhomogeneous removal rates. *Mathematical Biology*, 55:223–247, 2007.

T. Britton and N. G. Becker. Estimating the immunity coverage required to prevent epidemics in a community of households. *Biometrics*, 1(4):389–402, 2000.

T. Britton and P. D. O'Neill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3): 375–390, 2002.

T. Britton, T. Kypraios, and P. D. O'Neill. Inference for Epidemics with Three Levels of Mixing: Methodology and Application to a Measles Outbreak. *Scandinavian Journal of Statistics*, 38:578–599, 2011.

S. P. Brooks. Markov Chain Monte Carlo Method and its Application. *The Statistician*, 47:69–100, 1998.

A. Caimo and N. Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55, 2011.

J. Cardy and P. Grassberger. Epidemic models and percolation. *J. Phys. A*, 18(6): L267, 1985.

I. Chis-Ster and N. M. Ferguson. Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS ONE*, 2(6):e502, 2007.

C. Christensen, G. Bizhani, S-W. Son, M. Paczuski, and P. Grassberger. Agglomerative Percolation in Two Dimensions. *Europhys. Lett.*, 97:16004, 2012.

D. Clancy and P.D. O'Neill. Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, 3:737–758, 2008.

J. A. Clarkson and P. E. M. Fine. The efficiency of measles and pertussis notification in England and Wales. *International Journal of Epidemiology*, 14:153–168, 1985.

A. Cook, G. Marion, A. Butler, and G. Gibson. Bayesian inference for the spatio-temporal invasion of alien species. *Bulletin of Mathematical Biology*, 69:2005–2025, 2007.

D. R. de Souza and T. Tomé. Stochastic lattice gas model describing the dynamics of the SIRS epidemic process. *Physica A*, 389(5):1142–1150, 2010.

N. Demiris. *Bayesian Inference for Stochastic Epidemic Models using Markov chain Monte Carlo Methods.* PhD, University of Nottingham, 2004.

N. Demiris and P. D. O'Neill. Bayesian inference for epidemics with two levels of mixing. *Scand. J. Statist.*, 32(2):265–280, 2005.

N. Demiris and P. D. O'Neill. Computation of final outcome probabilities for the generalised stochastic epidemic. *Statistics and Computing*, 16(3):309–317, 2006.

O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computational ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.*, 28:365–382, 1990.

I. Dorigatti, S. Cauchemez, A. Pugliese, and N. M. Ferguson. A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: Application to the Italian 20092010 A/H1N1 influenza pandemic. *Epidemics*, 4(1): 9–21, 2012.

K. Eames and M. Keeling. Contact tracing and disease control. *Proc Biol Sci.*, 270 (1533):2565–2571, 2003.

N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, and S. Leach. Planning for smallpox outbreaks. *Nature*, 425:681–685, 2003.

J. Filipe, W. Otten, G. Gibson, and C. Gilligan. Inferring the dynamics of a spatial epidemic from time-series data. *Bulletin of Mathematical Biology*, 66:373–391, 2009.

B. F. Finkenstadt and B. T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Appl. Statist.*, 49, part 2:187–205, 2000.

C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. Ma. Espejo Guevara, F. Checchi, E. Garcia, S. Hugonnet, and C. Roth. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324: 1557–1561, 2009.

D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC Press, 2000.

C. Gerardo, N. Hiroshi, and L. M. Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J. R. Soc. Interface*, 4:155–166, 2007.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calcualting posterior moments. In *J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), Bayesian Statistics 4*, pages 169–193. Oxford: Oxford University Press, 1992.

G. Gibson. Markov chain monte carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Applied Statistics*, 46(2):215–233, 1997.

G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics Applied in Medicine and Biology*, 15:19–40, 1998.

G. J. Gibson, W. Otten, J. A. N. Filipe, A. Cook, G. Marion, and C. A. Gilligan. Bayesian estimation for percolation models of disease spread in plant populations. *Statistics & Computing*, 16:391–402, 2006.

G.J. Gibson, C.A.Gilligan, and A. Kleczkowski. Predicting variability in biological control of a plant-pathogen system using stochastic models. *Proc. Roy. Soc.*, 266: 1743–1753, 1999.

G.J. Gibson, A. Kleczkowski, and C.A.Gilligan. Bayesian analysis of botanic epidemics using stochastic compartmental models. *Proc. National Academy Sciences USA*, 101:12120–12124, 2004.

G.J. Gibson, G. Streftaris, and S. Zachary. Generalise data augmentation and posterior inferences. *J. Statist. Plan. Infer.*, 141:156–171, 2011.

W. R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81 (25):2340–2361, 1977.

P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983.

P. J. Green. Reversible jump MCMC computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

P. J. Green and D. I. Hastie. Reversible jump MCMC. Technical report, 2009.

C. M. Grinstead and J. L. Snell. *Introduction to Probability*. American Mathematical Society Providence RI, 1997.

W. K. Hastings. Monte Carlo sampling using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Y. Hayakawa, P.D. O'Neill, D. Upton, and P.S.F. Yip. Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Aust. N. Z. J. Stat.*, 45(4):491–502, 2003.

J. A. P. Heesterbeek and K. Dietz. The concept of R_0 in epidemic theory. *Statist. Neerlandica*, 50(1):89–110, 1996.

N. Hens, G. M. Ayele, N. Goeyvaerts, M. Aerts, J. Mossong, J. W. Edmunds, and P. Beutels. Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infectious Diseases*, 9, 2009.

N. Hens, M. Van Ranst, M. Aerts, E. Robesyn, P. Van Damme, and P. Beutels. Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: a multi-country analysis for influenza A/H1N1v 2009. *Vaccine*, 29:896–904, 2011.

H. W. Hethcote. The mathematics of infectious diseases. *SIAM Rev*, 42:599–653, 2000.

T. House and M. J. Keeling. The impact of contact tracing in clustered populations. *Computational Biology*, 6(3):e1000721, 2010.

V. Isham. Stochastic models for epidemics. In *Celebrating Statistics: Papers in Honour of Sir David Cox on His 80th Birthday*, pages 1–31. Oxford Statistical Science Series, 2005.

C. P. Jewell, T. Kypraios, P. Neal, and G. O. Roberts. Bayesian analysis for emerging infectious disease. *Bayesian Analysis 4*, 4:465–496, 2009.

C.P. Jewell, M.J. Keeling, and G.O. Roberts. Predicting undetected infections during the 2007 foot and mouth disease outbreak. *JRS Interface*, 6:1145–1151, 2008.

J. H. Jones. Notes on R_0, 2006. Available from `http://www.stanford.edu/~jhj1/teachingdocs/Jones-on-R0.pdf`.

E. H. Kaplan, D. L. Craft, and L. M. Wein. Emergency response to a smallpox attack: The case for mass vaccination. In *Proc. Nat. Acad. Sci*, volume 99, pages 10935–10940. PNAS, 2002.

M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton University Press, 2007.

W. O. Kermack and A. G. McKendrick. Contributions to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Part I*, A115:700–721, 1927.

J. F. C. Kingman. *Poisson Processes*. Oxford Studies in Probability, 3. The Clarendon Press, Oxford University Press, New York, 1993.

J. S. Koopman, C. P. Simon, J. A. Jacquez, and T. S. Park. Selective contact within structured mixing with an application to HIV transmission risk from oral and anal sex. In *Mathematical and statistical approaches to AIDS epidemiology VOLUME 83 of Lecture Notes in Biomath.*, pages 316–348. Springer, Berlin, 1989.

K. Kuulasmaa. The spatial general epidemic and locally dependent random graphs. *J. Appl. Prob.*, 19:745–758, 1982.

K. Kuulasmaa and D. Mollison. Spatial epidemic models: theory and simulations. *In: Bacon PJ, ed. The Population dynamics of rabies in wildlife. London: Academic Press*, pages 291–309, 1985.

K. Kuulasmaa and S. Zachary. On spatial general epidemics and bond percolation processes. *J. Appl. Prob.*, 21:911–914, 1984.

T. Kypraios. *Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and A New Class of SemiParametric Time Series Models.* PhD, Lancaster University, 2007.

T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes.* Springer, New York, 1999.

M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray. Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300:1966–1970, 2003.

I. M. Longini and J. S. Koopman. Household and community transmission parameters from final distributions of infections in households. *Biometrics*, 38(1):115–126, 1982.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

K. Mackenzie and S.C. Bishop. Developing stochastic epidemiological models to quantify the dynamics of infectious diseases in domestic livestock. *J. Anim. Sci.*, 79: 2047–2056, 2001.

G. Marion, G. J. Gibson, and E. Renshaw. Estimating likelihoods for spatio-temporal models using importance sampling. *Statistics and Computing*, 13:111–119, 2003.

N. Metropolis, A. W Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

D. Mollison. Spatial contact models for ecological and epidemic spread. *J. R. Statist. Soc. B.*, 39:283–326, 1977.

D. Mollison. *Epidemic Models: Their Structure and Relation to Data.* Cambridge University Press, Cambridge, 1995.

C.N. Morris. Parametric Empirical Bayes Inference: Theory and Applications. *J. Am. Stat. Ass.*, 78:47–55, 1983.

J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G.S. Tomba, J. Wallinga, J.Heijne, M. Sadkowska-Todys, M. Rosinska, and J. W. Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5, 2008.

P. Neal. Efficient likelihood-free bayesian computation for household epidemics. *Statistics and Computing*, (1997), 2010. URL `http://www.springerlink.com/index/10.1007/s11222-010-9216-x`.

P. Neal and G. Roberts. A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15:315–327, 2005.

J. R. Norris. *Markov Chains.* University of Cambridge Press, 1998.

B. K. Øksendal. *Stochastic Differential Equations: An introduction with applications.* Berlin: Springer, 2003.

P. D. O'Neill and N. Demiris. Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal Of The Royal Statistical Society Series B*, 67(5):731–745, 2005.

P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. C*, 49(4):517–542, 2000.

P.D. O'Neill and N.G. Becker. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2, 1:99–108, 2001.

P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. *J.R. Statist. Soc. A*, 162, Part 1:121–129, 1999.

O. Papaspiliopoulos, G. O. Roberts, and M. Skold. Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics, 7 (Tenerife 2002)*, pages 307–326, 2003.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006.

A. E. Raftery and S. M. Lewis. How Many Iterations in the Gibbs Sampler? In *J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), Bayesian Statistics 4*, pages 763–773. Oxford: Oxford University Press, 1992.

E. Renshaw. *Modelling Biological Populations in Space and Time*. Cambridge University Press, 1993.

W. N. Rida. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. Roy. Statist. Soc. Ser. B*, 53(1):269–283, 1991.

S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hedley, G. M. Leung, L.-M. Ho, T.-H. Lam, T. Q. Thach, P. Chau, K.-P. Chan, S.-V. Lo, P.-Y. Leung, T. Tsang, W. Ho, K.-H. Lee, E. M. C. Lau, N. M. Ferguson, and R. M. Anderson. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions. *Science*, 425:1961–1966, 2003.

G. O. Roberts and S. K. Sahu. Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. Roy. Statist. Soc. Ser. B*, 59(2): 291–317, 1997.

L. M. Sander, C. P. Warren, and I. M. Sokolov. Epidemics, disorder, and percolation. *Physica A: Statistical Mechanics and its Applications*, 325(1–2):1–8, 2003.

T. Sellke. On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Prob*, 20:390–394, 1983.

B. J. Smith. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software*, 21(11), 2007.

G. Streftaris and G. Gibson. Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, 4:63–75, 2004a.

G. Streftaris and G.J. Gibson. Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proc. R. Soc. Lond. B*, 271:1111–1117, 2004b.

G. Streftaris and G.J. Gibson. Non-exponential tolerance to infection in epidemic systems – modelling, inference and assessment. *Biostatistics*, pages 1–14, 2012. doi: 10.1093/biostatistics/kxs011.

G. Streftaris and B. J. Worton. Efficient and accurate approximate bayesian inference with an application to insurance data. *Computational Statistics & Data Analysis*, 52:2604–2622, 2008.

L. F. White and M. Pagano. Reporting errors in infectious disease outbreaks, with an application to pandemic influenza A/H1N1. *Epidemiologic Perspectives & Innovations*, 7, 2010.

P. Whittle. The outcome of a stochastic epidemic-a note on Bailey's paper. *Biometrika*, 42:116–122, 1955.

T. Williams. An algebraic proof of the threshold theorem for the general stochastic epidemic (abstract). *Adv. Appl. Prob*, 3:223, 1971.

S. Zachary. Bayesian inference and computational methods. Available from `http://www.ma.hw.ac.uk/~stan/f73bi/`, 2008.